RICE UNIVERSITY

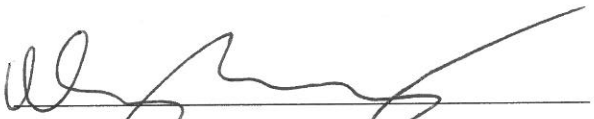
Statistical Modeling for Cellular Heterogeneity Problems in Cancer Research: Deconvolution, Gaussian Graphical Models and Logistic Regression

by

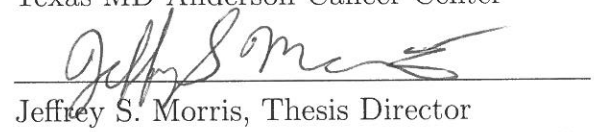
Zeya Wang

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree
Doctor of Philosophy

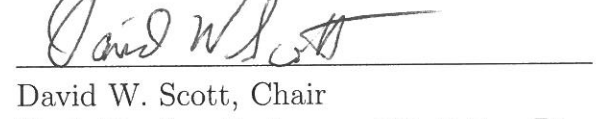
Approved, Thesis Committee:



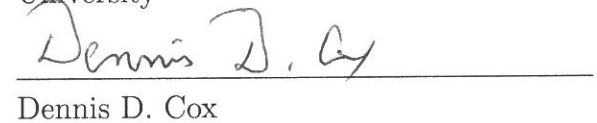
Wenyi Wang, Thesis Director
Associate Professor of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center



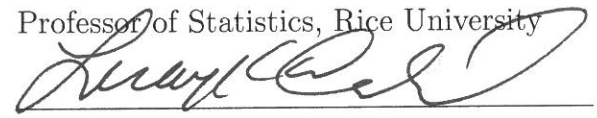
Jeffrey S. Morris, Thesis Director
Del and Dennis McCarthy Distinguished Professor of Biostatistics, The University of Texas MD Anderson Cancer Center



David W. Scott, Chair
Noah Harding Professor of Statistics, Rice University



Dennis D. Cox
Professor of Statistics, Rice University



Luay K. Nakhleh
Professor of Computer Science and of BioSciences, Rice University

Houston, Texas
April, 2017

ABSTRACT

Statistical Modeling for Cellular Heterogeneity Problems in Cancer Research: Deconvolution, Gaussian Graphical Models and Logistic Regression

by

Zeya Wang

Tumor tissue samples comprise a mixture of cancerous and surrounding normal cells. Investigating cellular heterogeneity in tumors is crucial to genomic analyses associated with cancer prognosis and treatment decisions, where the contamination of non-cancerous cells may substantially affect gene expression profiling in clinically derived malignant tumor samples. For this purpose, we first computationally purify tumor profiles, and then develop new statistical modeling techniques to incorporate tumor purity estimates for genetic correlation and prediction of clinical outcome in cancer research. In this thesis, we propose novel approaches to analyzing and modeling cellular heterogeneity problems using genomic data from three perspectives. First, we develop a computation tool, *DeMixT*, which applies a deconvolution algorithm to explicitly account for at most three cellular components associated with cancer. Compared with the experimental approach to isolate single cells, *in silico* dissection of tumor samples is faster and cheaper, but computational tools previously developed have limited ability to estimate cellular proportions and tumor-specific expression profiles, when neither is given with prior information. Our model allows inclusion of the infiltrating immune cells as a component as well as the tumor cells and stromal cells. We assume a linear mixture of gene expression profiles for

each component satisfying a log2-normal distribution and propose an iterated conditional modes algorithm to estimate parameters. We also involve a novel two-stage estimation procedure for the three-component deconvolution. Our method is computationally feasible and yields accurate estimates through simulations and real data analyses. The estimated cellular proportions and purified expression profiles can provide deeper insight for cancer biomarker studies. Second, we propose a novel edge regression model for undirected graphs, which incorporates subject-level covariates to estimate the conditional dependencies. Current work for constructing graphical models for multivariate data does not take into account the subject specific information, which can bias the conditional independence structure in heterogeneous data. Especially for tumor samples with inherent contamination from normal cells, ignoring the cellular heterogeneity and modeling the population-level genomic graphs may inhibit the discovery of the true tumor graph, which would be attenuated towards the normal graph. Our model allows undirected networks to vary with the exogenous covariates and is able to borrow strength from different related graphs for estimating more robust covariate-specific graphs. Bayesian shrinkage algorithms are presented to efficiently estimate and induce sparsity for generating subject-level graphs. We demonstrate the good performance of our method through simulation studies and apply our method to cytokine measurements from blood plasma samples from hepatocellular carcinoma (HCC) patients and normal controls. Third, we build a model with respect to logistic regression that includes tumor purity as a scaling factor to improve model robustness for the purpose of both estimation and prediction. Penalized logistic regression is used to identify variables (genes) and predict clinical status with binary outcomes that are associated with cancers in high-dimensional genomic data. We aim to reduce the uncertainty introduced by cellular heterogeneity through incorporating the measure of tumor purity to quantify the power of data for each sample. We provide strategies of choosing scaling parameters. Our model is finally shown to work well through a set of simulation studies. We believe that the statis-

tical modeling, technical pipelines and computational results included in our work will serve as a first guide for the development of statistical methods accounting for cellular heterogeneity in cancer research.

Acknowledgments

I feel truly grateful for all the people who helped me during my Ph.D. study and accompanied me through my five-year graduate study at Rice. I feel greatly fortunate for having their supports to my academic and daily life. My study and research at Rice University would not have been completed without their favors.

First and foremost, I would like to express my sincere gratitude to my two advisors, Prof. Wenyi Wang and Prof. Jeffrey S. Morris, who introduced me to the field of Biostatistics, and provided me with continuous financial support to finish my Ph.D. study and related research. As my main advisor, Prof. Wenyi Wang led me to start my research at MD Anderson Caner Center and provided me with this valuable opportunity of working in her research group for four years. Her broad knowledge and in-depth understanding in Biostatistics and Bioinformatics inspired me a lot and motivated me deeply for my research. I also thank her so much for her great patience and encouragement to provide accumulated experience and knowledge that helps me grow up. I am also very grateful to Prof. Jeffrey S. Morris for his constant support and valuable guidance for promoting my growth. His great wisdom and extensive knowledge in the field of Statistics benefited my work significantly.

Besides my advisors, I would like to give many thanks to my thesis committee members for their time, efforts, and guidance: Prof. David W. Scott, Prof. Dennis D. Cox and Prof. Luay K. Nakhleh. I deeply appreciate their valuable helps to me for proposing and accomplishing my thesis work. I would also like to thank many professors who fueled up my academic knowledge and built me a solid base of statistical skills. Their super talented sense on science navigate my journey to embrace the beauty of Statistics.

I acknowledge many professors and colleagues who have given me many helps in shaping my research work. These include but are not limited to Prof. Chris C. Holmes at the University of Oxford, Prof. Veerabhadran Baladandayuthapani and Prof. Hongtu Zhu at the University of Texas MD Anderson Cancer Center. I also much appreciate the great supports given by my group members and their company through those years, especially Shaolong Cao, Xuedong Pan, Jialu Li, Elissa Dodd and Rongjie Liu. I also want to give my special thanks to those collaborators involved in my thesis work. I feel so honored to have your precious support to move my projects forward.

I also met many wonderful friends who provide a lot of helps for my campus life. I want to thank Yang Ni, Quan Zhou, Shiting Lan, who ever provided helpful discussions for my research, and all those my friends with whom I had a great time. A special thank should be given to my host family in Houston, who always bring warmness to my after school life.

Last but not the least, I would give many great thanks to my family, especially my parents. They are the greatest treasure of my life and always stand behind me, encouraging me to overcome any difficulties and keep moving forward. Thank you both for your great love and support through my life.

Contents

Illustrations

Tables

Chapter 1

Introduction

Cellular heterogeneity has been documented as a crucial factor in cancer prognosis or treatment for many years. In genomic analyses, expression profiles observed from tumor samples can be contaminated and even dominated by non-cancer cells (normal cells). In this thesis, we propose three topics by applying different statistical modeling techniques from different perspectives for cellular heterogeneity problems in cancer research. The first topic involves *in silico* dissection of tumor samples using expression data. The second topic focuses on more general covariate-dependent undirected graphical models while accounting for biological heterogeneity in construction of regulatory networks. The third topic deals with penalized logistic regression by introducing heterogeneity to classification of binary outcomes. This chapter introduces our motivation and contributions as well as the background and problem. Section 1.1 gives a description of the well-defined problems for cellular heterogeneity. Section 1.2 explains our scientific motivations for this problem. Section 1.3 provides an outline of the dissertation and an overview of our methods. Section 1.4 describes our contributions.

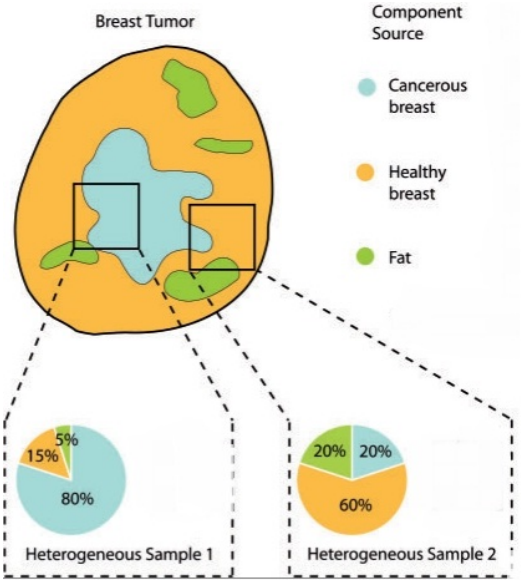
1.1 Cellular Heterogeneity of Tumors

Tumor is not just one whole blob (Fig. 1.1). Solid tumors are heterogeneous from a morphological prospective. Tumors are neoplasm that grow within an intricate en-

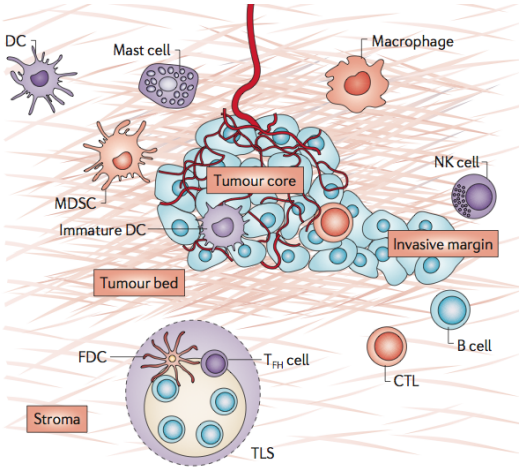
vironment composed of different types of cells (Marusyk et al., 2012). Cancer cells "huddle" with a highly diverse cell populations, including epithelial cells, mesenchymal cells, endovascular cells, cytokines and chemokines, and infiltrating immune cells (Marusyk et al., 2012). This very complex tissue architecture of tumors defines the cellular heterogeneity problem (Gay et al., 2016; Sun and Yu, 2015; Marjanovic et al., 2013; Meacham and Morrison, 2013; Marusyk et al., 2012; Xu et al., 2014). Actually, each population is also heterogeneous as a mixture of different cells. Those non-cancerous cells are from surrounding healthy tissues, which are interacted with cancer stem cells during the tumor growth and disease progression, including localization, growth, invasion, extravasation, therapy resistance and metastasis (Pages et al., 2009). Not only the cancer cells, but also those cells involved will determine the progression of cancer, and cause molecular differences in tumors. Those differences affect clinical diagnostics and therapeutic response, which can cause dramatical variations even for cancer patients with very similar clinical characteristics. Tumor purity, which is the percentage of cancerous cells and also defines the degree about how tumor genome sequence differ from the normal's after genetic mutation, can vary widely among samples, even with a high cancerous content for the same cancer type (Su et al., 2012; Quon et al., 2013).

Gene/protein expression profiling are measurements of mRNA/active protein level to provide a high-resolution picture of molecular functions (Liotta and Petricoin, 2000). They determine the pattern of gene expression, and can be used to detect the underlying transcriptional activity and reveal its aberrant behaviors (Liotta and Petricoin, 2000; Quon and Morris, 2009). A lot of statistical methods based on gene/protein expression profiles are developed for genomic analyses in cancer research, but a majority of current analysis methods treat cancer as homogeneous and do not account for

this aspect of cellular heterogeneity, which introduces variability/noises into expression profiles (Aran et al., 2015). This variability is hard to be removed by available analytic methods and may confuse the detection of gene signatures associated with cancer prognosis. Furthermore, those confounding transcriptional signals brought by non-cancerous cells will reduce the effective sample size and thus the statistical power for genomic studies in cancer (Quon et al., 2013). It has been widely believed that the utility of profiles with common analytical techniques is limited by cellular heterogeneity in tumors (de Ridder et al., 2005).



(a) Tumor comprises cancerous tissues and healthy tissues. Multiple samples extracted from the same tumor have different proportions of constituent tissues and exhibit different expression profiles. (Figure source: (Quon and Morris, 2009))



(b) Tumor core is interwoven with immune contexture (the invasive margin, tertiary lymphoid structures, et.al.) and tumor microenvironment (stroma, et.al). (Figure source: (Fridman et al., 2012))

Figure 1.1 : Demonstration for cellular heterogeneity in tumors.

1.2 Scientific Motivation

Our works, in this thesis, stem from the fact that cancer is heterogeneous at the cellular and molecular levels. Cellular heterogeneity problems may significantly bias analyses based on gene expression profiles and lead to false conclusions (Farley, 2015). Statistical modeling and methods are important tools used to find biomarkers for cancer prognosis and treatment. But this cellular heterogeneity problem has not been thoroughly investigated, and thus can bring those negative effects mentioned previously. As a consequence, extracting tumor purity from mixed tumor samples and modeling compartment-specific information are required for us to seek for significant improvements in cancer research.

We are first motivated to deconvolve the observed expression profiles of tumor samples into those contributed by different cell sub-populations/types (Xu et al., 2014). We believe that for a set of cell type-specific genes, expression profiles uniquely reflect the functional characteristics of the corresponding cell types. A deconvolution (deconvolution) framework is developed to decompose the expression dataset collected from tumor samples, which are assumed as a mixture of different cell types, into expression profiles specific to each component. This is our first step, but also the key step. After deconvolution, we obtain the proportions of assumed cancerous component, which correspond to the tumor purity mentioned above. We first begin with the assumption that a solid tumor sample is a mixture of two distinct components, cancerous and non-cancerous (e.g., normal), which has been frequently assumed in other literature and works (Ahn et al., 2013; Wang et al., 2015). This simple assumption allows deconvolution to be performed under a clearer context, but we need to overcome several challenges for deconvolution technique. A more relaxed model is expected to include more subdivided components, when we realize the immune contexture is a crucial

factor on tumor (Fridman et al., 2012). A better deconvolution algorithm helps us to get better estimates of tumor purity and model expressions contributed by separate cell types more accurately.

Following the work of tumor deconvolution, we investigate how to incorporate tumor purity estimates into statistical modeling for genetic correlation and prediction of clinical outcome, which are two of the most common approaches used in cancer research. Gaussian graphical models are usually applied, when gene regulatory networks are constructed to model the regulatory relationships among genes and their products. Tumor purity may introduce nontrivial bias into the constructed networks, when a lot of current works consider tumor samples as homogeneous following our previous discussion. More interestingly, the variability brought by difference of tumor purity for patient samples can give rise to topological change in the regulatory networks, which drive the progression of cancer, at an individual level. Thus, we want to include a "personalized" network instead of a general network for all those patients (Ni et al., 2017). This work can be complicated because we need to maintain the "personalized" conditional dependency structure among genes for all the individual samples. More generally speaking, current statistical work for constructing graphical models for multivariate data does not take into account the subject specific information. We believe this can inhibit the discovery of the true conditional dependency structure in the heterogeneous data. We are motivated to develop a more general statistical framework for subject-varying graphical models. Diagnostic classification of clinical status for patients and finding predictive genes for case-control disease is an integral part in cancer research. Logistic regression is a standard method used to implement binary classification and feature selection with penalized likelihood in genomic studies. The contamination of normal cells and the measurement error of

deconvolved expression data can mislead the prediction of clinical outcome mainly associated with cancerous component. Therefore, we want to elicit a model to quantify and introduce the uncertainty caused by tumor purity into the logistic regression for cancer data.

1.3 Outline of the Dissertation

In this dissertation, we aim at different statistical methods for the cellular heterogeneity problems mainly from three prospective, including deconvolution, Gaussian graphical models and logistic regression. The goal of our modeling and analysis approach is to improve the power of conventional statistical tools in cancer research, when the heterogeneity in data undermines their assumptions. In each different work, we first extract the cancer specific information from the clinically derived malignant tumor samples and integrate this subject-level information into the downstream biomarker studies. We build and test our models on simulated data and real data, which include microarray expression data, RNA-seq expression data and even cytokine expression data. The focus of this dissertation is the performance of our method in each section and biological findings we discover by applying our methods.

The first chapter of this thesis is an introductory overview of the cellular heterogeneity problems in cancer research and its background information. We explain the statistical and biological significance of addressing those problems and take a review of the challenges we are facing. We also point out contributions of our work as well as implications for the future studies.

The following three chapters address three topics: deconvolution, Gaussian graphical models and logistic regression, for cellular heterogeneity problem. Chapter 2 describes

the work on the transcriptome deconvolution of mixed tumor samples with immune infiltration. Deconvolution, which is the process of resolving the observed gene expression data into expressions contributed by different components we assume for different cell types, is realized by finding an appropriate derivative-free optimization procedure for a log-normal distribution based model. We begin to model two separate components for cancerous and non-cancerous tissues and extend to three components in our model first to accommodate the existence of infiltrating immune cells. We develop an *R* package *DeMixT* to incorporate all those features we build for tumor deconvolution and make it to be implemented fast and easily. We demonstrate results for estimating component-specific proportions and deconvolved expressions through several computationally and biologically simulated datasets, and one real biological data set. We also reveals several interesting biological findings after applying our deconvolution tool.

The next two chapters investigate biomarker studies for tumor samples from the perspective of constructing gene regulatory networks and performing diagnostic prediction. Chapter 3 presents the work on the Bayesian edge regression. We introduce a novel class of Bayesian edge selection model to allow the topology of high-dimensional undirected graphs dependent on exogenous covariates in a flexible way. We define our edge regression model for undirected graphs and provide some theoretical properties of our model. We propose a joint regression model with a Bayesian inference approach to preserve the symmetry among the partial correlations, which is ignored by all those methods developed previously for *conditional covariance selection*. We validate our model in two cases for continuous and categorical exogenous covariates by comparing with several competitive methods. Specifically, the case for continuous covariates is simulated corresponding to the assumed model in the tumor deconvolution problem,

so that we can employ it directly to estimate varying structure networks for tumor samples with different tumor purity. Finally, we illustrate the application of our edge regression model in a liver cancer cytokine study to estimate blood plasma cytokine networks induced by hepatocellular carcinoma and those from normal controls while accounting for biological heterogeneity. In Chapter 4, we develop a logistic regression model with a scaling factor function, which can be also applied in a penalized form. For diagnostic prediction for tumor samples, we consider observations with different tumor purity to contribute differently to predicting binary outcomes associated with tumors. Our model allows tumor purity to control the mean of the Bernoulli distribution through linking it with the scaling parameter in the sigmoid function. We present how to realize it in the logistic regression with and without penalty term, and finally validate our model through the simulation study.

Finally, Chapter 5 is a concluding section to summarize our work. We also point out several aspects from our current work, which can be extended and would become interesting development in the future for study of cellular heterogeneity problem.

1.4 Contribution

As discussed in the previous section, statistical modeling for cellular heterogeneity should necessarily soon be built into cancer research in the clinical and laboratory. Tumor purity cannot be a neglected factor in any genomic studies based on tumor samples. Existing methods that primarily focus on tumor deconvolution problem still have limited utility, and there are few statistical methods developed to consider the intrinsic heterogeneity of tumor samples in the downstream analysis for biomarker studies. The contributions of this dissertation are as follows. First, we propose a sta-

tistical method to jointly estimate tumor proportions and cancerous expressions when neither is given with prior knowledge. Our method is able to accurately deconvolve tumor samples into two or three cell subpopulations, which accommodate infiltrating immune cells as a single component, and allow for more variations on the referenced normal components. Second, we develop an *R*-package to integrate all those features we propose for deconvolution into standard *R*-based analysis pipeline. Our package is user friendly and easy for testing and modifying. Third, we begin to talk about the identifiability problem in deconvolution, which has been ignored in all the previous works and provide several suggestions in our discussion. Moreover, being motivated by how tumor purity affects co-expression network, we introduce a new class of undirected graphical model, Bayesian edge regression, to allow the undirected network structure vary with additional subject-level covariates and borrow strength from different related graphs for estimating more robust covariate-specific graphs. We further apply our novel edge regression model to the cellular heterogeneity problem, to infer more robust tumor and normal graph as well as subject-level graphs, for different cancer samples, therefore facilitate more meaningful biological interpretation. Finally, we incorporate tumor purity into classification of binary outcomes for genome-wide association study to improve the prediction performance.

Chapter 2

Cell type-specific Deconvolution of Heterogeneous Tumor Samples using Expression Data

2.1 Abstract

Following the discussion above, it is crucial to analyze gene signatures associated with cancer prognosis and treatment decisions by investigating their cellular heterogeneity. To begin with this study, we first need to deconvolve observed expression data of tumor samples. Compared with the experimental approach of laser micro-dissection to isolate different tissue components (Emmert-Buck et al., 1996), *in silico* dissection of mixed cell samples, which is enabled through computational tools, is faster and cheaper. Computational approaches previously developed have their limitations to deconvolve tumor profiles. We have developed a deconvolution model, *DeMixT*, which has been integrated to the *R* package *DeMixT*, that can explicitly account for at most three components of tumor mixtures. Our method is able to address this challenging problem when the observed signals in tumor profiles are assumed to come from a mixture of cancerous tissues, infiltrating immune cells and tumor microenvironment or a mixture of just cancerous and non-cancerous tissues. *DeMixT* is computationally feasible when it is needed to compute high-dimensional integrals, and involves a novel two-stage filtering method that yields accurate estimates of cell type-specific proportions and compartment-specific expression profiles. Simulations and real data analyses have demonstrated the good performance of our method. *DeMixT* allows

for a further understanding of cellular heterogeneity in cancer, therefore assists the development of novel prognostic markers and therapeutic strategies.

2.2 Introduction

Heterogeneity of malignant tumor cells adds confounding complexity to cancer treatment. From carcinoma '*in situ*', tumor interacts with its microenvironment, which are comprised of non-cancerous stromal and immune cells, during the cancer progression stages of localization, growth, invasion, extravasation and metastasis (Kalluri and Zeisberg, 2006; Pages et al., 2009; Fridman et al., 2012).The evaluation of individual component of tumor samples is complicated by tumor-stroma interaction and tumor infiltration by lymphocytes, macrophages or mast cells. Experimental approaches, such as laser micro-dissection and cell sorting, are limited by their economic and time costs. The vast expression datasets, which are generated by gene expression profiling technique, motivate the development of computational tools to deconvolve mixed tumor samples on the expression levels (de Ridder et al., 2005; Shen-Orr and Gaujoux, 2013).

A majority of methods for *in silico* deconvolving tumor samples are developed to deconvolve transcriptomes by using microarray expression data and HTS data (Yadav and De, 2015). There are two challenges that lack to be solved by most available methods for deconvolution: (i) no violation of linear additive relationship of between tumor profiles and its constituent profiles; (ii) estimation of both cell type-specific proportion and tumor-specific expression for individual samples at the same time when neither is given with priori (Ahn et al., 2013). Several approaches require reference gene sets to be imported into their pipelines (Gong and Szustakowski, 2013; Ahn et al., 2013). Some method deconvolves two samples at a time, but is not able to

identify cancerous and non-cancerous components (Wang et al., 2015). Several methods are developed following a matched pattern design, which requires tumor samples and reference normal samples to be derived from the same individual (Quon et al., 2013; Ahn et al., 2013). However, the clinical routine that deconvolves tumor samples in an unmatched pattern limits the application of those computational tools.

Most commonly available deconvolution methods assume that malignant tumor cells consist of two distinct components, epithelium-derived tumor and surrounding stromal cells. However, immune infiltration to tumor cells is a crucial factor in cancer prognosis. Evidence from epidemiological studies suggests that chronic inflammation, which is initiated and stimulated by immune cells, promotes tumor growth (Pages et al., 2009). Anatomical studies of the tumor-immune cell contexture have demonstrated that it primarily consists of a tumor core, lymphocytes and the tumor microenvironment (Pages et al., 2009; Fridman et al., 2012). Further research supports the association of infiltrating immune cells with clinical outcome for individuals with ovarian cancer, colorectal cancer and follicular lymphoma (Zhang et al., 2003; Dave et al., 2004; Galon et al., 2006). Therefore, understanding the heterogeneity of tumor cells motivates a computational method to integrate reconstruction of expression profiles for immune cells in addition to the cancerous cells with its microenvironment. Other deconvolution methods for more than two compartments require knowledge of cell-component-specific gene lists (Liebner et al., 2014; Yoshihara et al., 2013), but do not provide joint estimates of tissue proportion and tissue-specific gene expression.

In this work, we develop a statistical approach, *DeMixT*, for deconvolution of gene expression data from mixed tumor samples that is able to account for at most three components. It differs from previous methods in its ability to estimate cell-type-specific proportions and then the full expression matrices for all assumed mixing

components (Fig. 2.1). We take advantage of normality of log2-transformed expression data as well as the linear addition of raw expression data from constituent tissues for tumor profiles (Ahn et al., 2013; Lönnstedt and Speed, 2002). We integrate our method into *DeMixR* pakcage as a new feature that can be applied to deconvolve tumor profiles into components from cancerous, stromal and immune cells. It is known that the immune component is heterogeneous, as its own composition varies in terms of immune cell types. However, it was reported that the level of heterogeneity within this component may be similar and consistent for its relative proportions of immune cell types (Gentles et al., 2015). Therefore, that is the reason we may be able to model expressions of tumor-infiltrating immune cells as a single component made up of a stable mixture of immune cell types. *DeMixT* finds a good local optimum by employing the method of iterated conditional modes that cyclically maximizes the probability of each variable conditionally on the rest, for which we have observed rapid convergence. *DeMixT* also utilizes a novel two-stage method to filter out reliable expression measurements to remove biological noise and improve estimation performance. In the next, we demonstrate the performance of *DeMixT* through simulation studies and real data analyses.

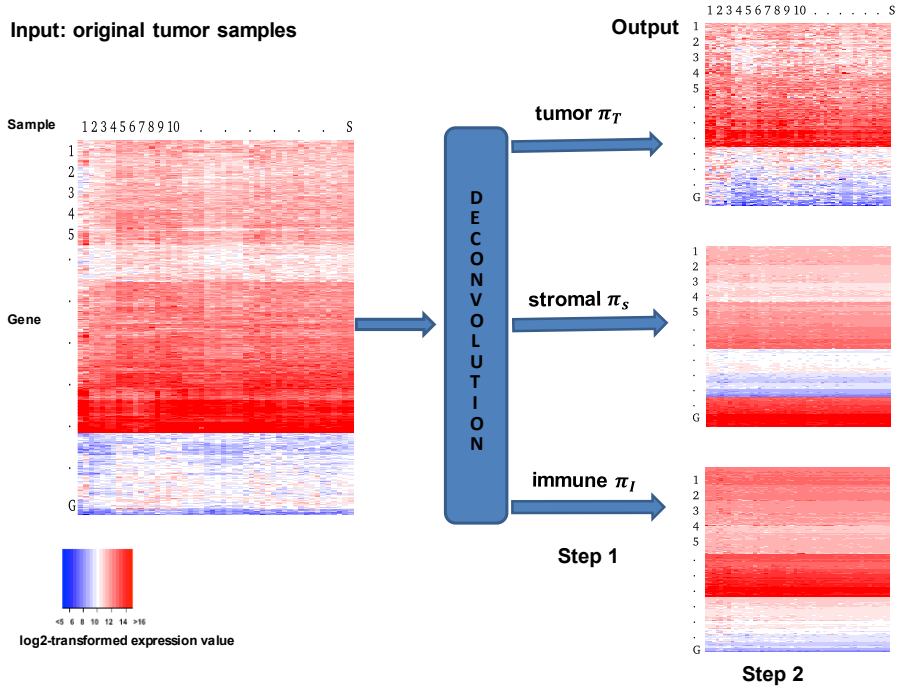


Figure 2.1 : Three-component deconvolution to output tissue-specific proportion, and isolated expression matrices of tumor, stromal and immune cells. Heat map of expression levels uncovers the difference in gene expression patterns between original tumor samples, deconvolved tumor components, stromal components and immune components

2.3 Methods

2.3.1 Deconvolution model

Let Y_{ig} be the observed expression levels of the raw measured data from clinically derived malignant tumor samples for gene $g, g = 1, \cdots, G$ and sample $i, i = 1, \cdots, S$. G denotes the total number of features (e.g., probes or genes) and S denotes the number of samples. The observed expression levels for solid tumors can be modeled as a linear mixture of raw expression levels from of its constituents (Ahn et al., 2013). The immune composition in bulk tumors is also heterogeneous, which complicates the model if all the specific tumor infiltrating immune cell subsets need to be included.

However, it was recently reported that given a cancer type, even when multiple subsets exist, the relative composition of leukocyte cells is consistent across different samples (Gentles et al., 2015). Therefore, for solid tumors of some specific cancer type, heterogeneous immune composition, considered as a consistent combination of different subsets, can be assumed as one mixing component in our model:

$$Y_{ig} = \pi_{1,i} N_{1,ig} + \pi_{2,i} N_{2,ig} + (1 - \pi_{1,i} - \pi_{2,i}) T_{ig} \tag{2.1}$$

Let $N_{1,ig}$, $N_{2,ig}$ and T_{ig} be the unobserved expression levels from different components. For simplicity, we call the first two components N_1-component and N_2-component, the distribution parameters of which can be estimated from reference profiles for each tissues type. Those two components can be reduced to one, generating a two-component deconvolution model that has often been assumed for the cancerous and non-cancerous tissue. We define the last component as T-component, which refers to the unknown component that is not given with any previous information. Our model is different from previous methods by allowing one component to be unknown so not requiring reference profiles from all the constituents. A set of unmatched observations for $N_{1,ig}, N_{2,ig}$ is provided for deconvolution. The source of these unmatched observations is from other patients or historical data. $\pi_{1,i}$ denotes the proportion of stromal cells (N_1-component) and $\pi_{2,i}$ denotes the proportion of immune cells (N_2-component). Thus, $1 - \pi_{1,i} - \pi_{2,i}$ is the proportion of tumor (T-component) cells. We assume that the mixing proportions of one specific sample are the same across all genes.

Following the convention that log2-transformed microarray gene expression data follow a normal distribution, we assume that the raw measures $N_{1,ig} \sim LN(\mu_{N_1 g}, \sigma^2_{N_1 g})$,

$N_{2,ig} \sim LN(\mu_{N_2g}, \sigma^2_{N_2g})$ and $T_{ig} \sim LN(\mu_{T_g}, \sigma^2_{T_g})$, where LN denotes a log2-normal distribution and $\sigma^2_{N_1g}$,$\sigma^2_{N_2g}$,$\sigma^2_{T_g}$ reflect the variations under log2-transformed data (Ahn et al., 2013; Lönnstedt and Speed, 2002).

This mixed model can be expressed as the convolution of the density function for three log2-normal distributions. Because there is no closed form of this convolution, numerical integration was used to evaluate the likelihood function

$$
\begin{aligned}
L &= \prod_{i=1}^{S}\prod_{g=1}^{G} f(y_{ig} \,|\mu_{T_g}, \mu_{N_1g}, \mu_{N_2g}, \mu_T\sigma_{N_1g}, \sigma_{N_2g}\ , \sigma_{T_g}, \pi_{1,i}, \pi_{2,i}) \\
&\propto \prod_{i=1}^{S}\prod_{g=1}^{G} \{\int_0^y \frac{1}{{n'}_{2,ig}\sigma_{N_2g}} \exp[-\frac{\{\log 2({n'}_{2,ig}) - \mu_{N_2g} - \log 2(\pi_{2,i})\}^2}{2\sigma^2_{N_2g}}] \\
&\times \int_0^{y-{n'}_2} \frac{1}{{n'}_{1,ig}\sigma_{N_1g}} \exp[-\frac{\{\log 2({n'}_{1,ig}) - \mu_{N_1g} - \log 2(\pi_{1,i})\}^2}{2\sigma^2_{N_1g}}] \frac{1}{(y_{ig} - {n'}_{1,ig})\sigma_{T_g}} \\
&\times \exp[-\frac{\{\log 2(y_{ig} - {n'}_{1,ig} - {n'}_{2,ig}) - \mu_{T_g} - \log 2(1 - \pi_{1,i} - \pi_{2,i})\}^2}{2\sigma^2_{T_g}}] d{n'}_{1,ig} d{n'}_{2,ig}\}
\end{aligned}
\tag{2.2}
$$

, where $n'_{1,ig} = \pi_{1,i}n_{1,ig}$ and $n'_{2,ig} = \pi_{2,i}n_{2,ig}$

2.3.2 Inference

Two-step approach

DeMixT estimates all parameters (including cellular proportions) in equation 2.2 and reconstitutes the expression profiles in two steps.

1. Obtain a set of parameters $\{\pi_{1,i}, \pi_{2,i}\}_{i=1}^{S}$, $\{\mu_T, \sigma_T\}_{g=1}^{G}$ to maximize the complete likelihood function after inferring $\{\mu_{N_{1,g}}, \sigma_{N_{1,g}}, \mu_{N_{2,g}}, \sigma_{N_{2,g}}\}_{g=1}^{G}$ from unmatched samples for N_1 and N_2-component tissues.

2. Reconstitute the expression profiles by searching each pair of $\{n_{1,ig}, n_{2,ig}\}$ that maximize the joint density of $N_{1,ig}$, $N_{2,ig}$ and T_{ig}

$$\arg\max_{n_{1,ig},n_{2,ig}} \phi(\frac{y_{ig} - \hat{\pi}_{1,i} n_{1,ig} - \hat{\pi}_{2,i} n_{2,ig}}{1 - \hat{\pi}_{1,i} - \hat{\pi}_{2,i}} \left| \hat{\mu}_{T_g}, \hat{\sigma}_{T_g} \right.) \phi(n_{1,ig} \left| \hat{\mu}_{N_{1g}}, \hat{\sigma}_{N_{1g}} \right.) \phi(n_{2,ig} \left| \hat{\mu}_{N_{2g}}, \hat{\sigma}_{N_{2g}} \right.) \tag{2.3}$$

where $\phi(.|\mu, \sigma^2)$ is a log2-normal distribution density with location parameter μ and scale parameter σ.

In step 1, the estimation of $\{\mu_T, \sigma_T\}_g$ for any given gene g in a subset can be made separately after estimating $\{\pi_{1,i}, \pi_{2,i}\}_{i=1}^S$ for all the samples, where we just use these most identifiable genes for estimation. By maximizing the likelihood function given $\{\hat{\pi}_{1,i}, \hat{\pi}_{2,i}\}_{i=1}^S$, $l(\{\mu_T, \sigma_T\}_g|, \{y_{ig}, \hat{\pi}_{1,i}, \hat{\pi}_{2,i}\}_{i=1}^S)$, for each individual gene g, we obtain the maximum likelihood estimator of $\{\mu_T, \sigma_T\}_g$ independently. We remove genes that are estimated with very large $\hat{\sigma}_{Tg}$, as we consider these genes to be estimated inaccurately. We then reconstitute tissue-specific expressions of each individual.

Iterated conditional modes

First, using reference samples of N_1 and N_2 tissue types, we estimated $\{\mu_{N_1}, \sigma_{N_1}, \mu_{N_2}, \sigma_{N_2}\}_{g=1}^G$ through the method of moments. We include these estimates in our objective likelihood function, then the remaining unknown parameters can be assigned to two groups: genome-wise parameters, $\{\mu_T, \sigma_T\}_{g=1}^G$, and sample-wise parameters, $\{\pi_1, \pi_2\}_{i=1}^S$. We consider the derivation of the maximum likelihood estimators of our complete likelihood function as a problem of maximum a posteriori probability (MAP) estimation with non-informative priors. Then by considering unknown pa-

rameters as a set of variables, we can use a simple directed acyclic graph to describe the dependency structure among all the variables in our model. By demoralizing this directed graph, we obtain a graph to describe the conditional dependencies among sample-wise and genome-wise parameters, as shown in (Fig. 2.2). For each pair of genome-wise parameters (considered as variables), we have

$\{\pi_1, \pi_2\}_i \perp\!\!\!\perp \{\pi_1, \pi_2\}_j \Big| \{\mu_T, \sigma_T\}_1, \cdots, \{\mu_T, \sigma_T\}_G, \{y_{ig}\}_{i,g}^{S,G}$, for all $i \neq j \in \{1, \cdots, S\}$ and similarly for each pair of sample-wise parameters (considered as variables), $\{\mu_T, \sigma_T\}_i \perp\!\!\!\perp \{\mu_T, \sigma_T\}_j \Big| \{\pi_1, \pi_2\}_1, \cdots, \{\pi_1, \pi_2\}_S \{y_{ig}\}_{i,g}^{S,G}$, for all $i \neq j \in \{1, \cdots, G\}$. These relationships motivate us to design an optimization method to iteratively derive the modes of joint density for each pair of genome-wise or sample-wise parameters, conditional on the rest (Besag, 1986).

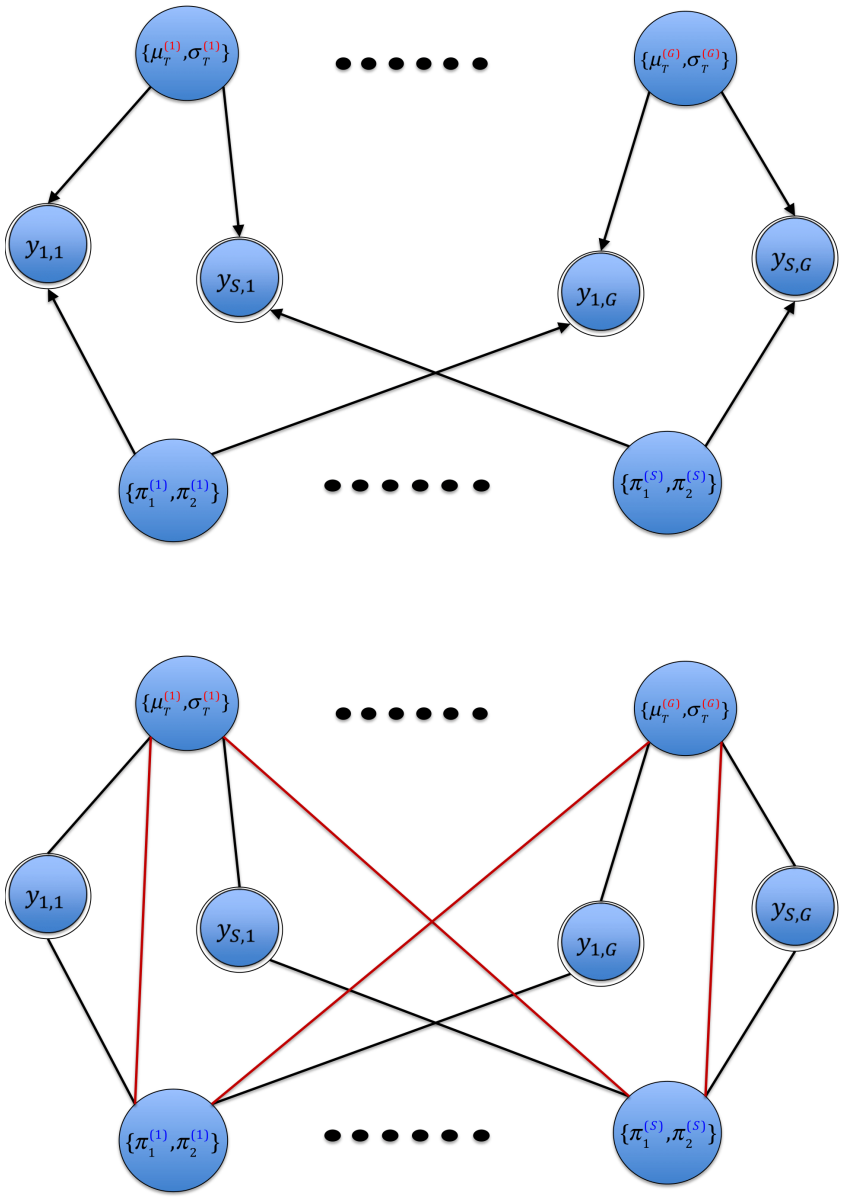


Figure 2.2 : Graphical representation of our underlying model. Nodes denote all the variables representing unknown parameters and observed expression profiles. They are connected with edges, which suggest conditional dependency structure.

According to our objective in principle, where π_1, π_2 are constrained between 0 and 1 and μ_T, σ_T are bounded positive, we combined a golden section search and

successive parabolic interpolations to find a good local maximum in each step (Brent, 1973). Our ICM procedure is organized as follows:

a. Initialize candidate $\{\mu_T^{(0)}, \sigma_T^{(0)}\}_{g=1}^G$

b. In each iteration $t = 1, \cdots, T$

• Step I:

Jointly search the pair $\{\pi_1^{(t)}, \pi_2^{(t)}\}_{i=1}^S$ to maximize the likelihood given $\{\mu_T^{(t-1)}, \sigma_T^{(t-1)}\}_{g=1}^G$ w.r.t

$$f(\{\pi_1, \pi_2\}_i \left| \{y_{ig}\}_{g=1}^G, \{\mu_T^{(t-1)}, \sigma_T^{(t-1)}\}_{g=1}^G\right.) = \prod_{g=1}^G f(y_{ig} \left| \{\pi_1, \pi_2\}_i, \{\mu_T^{(t-1)}, \sigma_T^{(t-1)}\}_{g=1}^G\right.), \quad \forall i = 1, \cdots, N \tag{2.4}$$

• Step II:

Jointly search the pair $\{\mu_T^{(t)}, \sigma_T^{(t)}\}_{g=1}^G$ to maximize the likelihood given $\{\pi_1^{(t)}, \pi_2^{(t)}\}_{i=1}^S$ w.r.t

$$f(\{\mu_T, \sigma_T\}_g \left| \{y_{ig}\}_{i=1}^S, \{\pi_1^{(t)}, \pi_2^{(t)}\}_{i=1}^S\right.) = \prod_{i=1}^S f(y_{ig} \left| \{\pi_1^{(t)}, \pi_2^{(t)}\}_{i=1}^S, \{\mu_T, \sigma_T\}_g\right.), \quad \forall g = 1, \cdots, G \tag{2.5}$$

c. Repeat each iterative steps until the convergence criteria is satisfied ($|L^t - L^{t-1}| < 10^{-5} \times L^{t-1}$).

In these iterative steps of ICM, the complete likelihood is updated by searching conditional modes. It never decreases at any iteration and the eventual convergence to the local maximum is guaranteed (Appendix A.1). To find a good local optimum, we employ random initializations of $\sigma_{Tg}^{(0)}$, and the initialization of $\mu_{Tg}^{(0)}$ is given by

$\log(3\bar{Y} - \exp(\mu_{N_{1g}}\log 2 + \sigma^2_{N_{1g}}(\log 2)^2) - \exp(\mu_{N_{2g}}\log 2 + \sigma^2_{N_{2g}}(\log 2)^2)) - (\sigma^{(0)}_{Tg})^2(\log 2)^2/2)\log(-2)$ (two-component: $\log(2\bar{Y} - \exp(\mu_{N_{1g}}\log 2 + \sigma^2_{N_{1g}}(\log 2)^2) - (\sigma^{(0)}_{Tg})^2(\log 2)^2/2)\log(-2))$. This is derived from $E(Y) = E(\pi_1)E(N_1) + E(\pi_2)E(N_2) + E(\pi_T)E(T)$ through assuming (π_1, π_2, π_T) satisfying Dirichlet distribution with equivalent concentration parameters.

Two-stage estimation

We develop a two-stage estimation procedure for the three-component deconvolution work. There are a great majority of genes in our test biological data sets with very similar expression levels across different cell types (Table 2.1). The selection of genes that have expression values observationally different across different components may challenge the inference. Differential expression analysis requires information of all those three components, but the T-component is blinded, i.e., we cannot observe expression patterns from reference profiles for T_{ig}. Also, traditional differential expression analysis is implemented between two components instead of three and cannot directly quantify the gene expression difference between different tissues. Thus, we come out of an intuitive measure for average expression difference among three components. Through assuming a Dirichlet distribution for proportions, i.e., $\{\pi_1, \pi_2, \pi_T\}_i \sim Dir(\alpha_1, \alpha_2, \alpha_3)$, where $\alpha_0 = \sum_{i=1}^{3} \alpha_i$. The variance for observed tumor profiles is given by (Goodman, 1960):

$$
\begin{aligned}
&Var(Y_g) = \\
&\frac{1}{\alpha_0^2(\alpha_0+1)}[\alpha_1\alpha_0(\alpha_1+1)Var(N_{1g}) + \alpha_2\alpha_0(\alpha_2+1)Var(N_{2g}) + \alpha_3\alpha_0(\alpha_3+1)Var(T_g) \\
&\quad + \alpha_1\alpha_2\{E(N_{1g}) - E(N_{2g})\}^2 + \alpha_1\alpha_3\{E(N_{1g}) - E(T_g)\}^2 + \alpha_3\alpha_2\{E(N_{2g}) - E(T_g)\}^2]
\end{aligned}
\tag{2.6}
$$

Table 2.1 : Number of genes with different relationships between different component tissues we summarize from datasets in our experimental validation and real data analysis; we defined $\mu_1 = \mu_2$ by satisfying $\frac{\mu_1}{\mu_2} < 1.1$ and $\frac{\mu_1}{\mu_2} > 0.9$.

Unknown Tissue	Number of Genes	Percentage of Genes
GEO19830:		
$\mu_{liver} = \mu_{brain} = \mu_{lung}$	19104/31099	61.1%
$\mu_{liver} \neq \mu_{brain} = \mu_{lung}$	2432/31099	7.8%
$\mu_{liver} = \mu_{brain} \neq \mu_{lung}$	1950/31099	6.3%
$\mu_{liver} \neq \mu_{brain} \neq \mu_{lung}$	1260/31099	4.1%
RNA-seq mixed cell line experiment:		
$\mu_{H1092} = \mu_{CAF} = \mu_{TIL}$	1503/5715	26.3%
$\mu_{H1092} \neq \mu_{CAF} = \mu_{TIL}$	628/5715	11.0%
$\mu_{H1092} = \mu_{CAF} \neq \mu_{TIL}$	979/5715	17.1%
$\mu_{H1092} \neq \mu_{CAF} \neq \mu_{TIL}$	1227/5715	21.5%
Laser capture microdissected prostate cancer patient samples:		
$\mu_{Tumor} = \mu_{Normal}$	32128/32321	99.4%
$\mu_{Tumor} \neq \mu_{Normal}$	193/32321	0.6%

Therefore, from the expression of variance for mixed tumor samples, there is a trade-off between the variance for individual cells and the difference of the average expression levels between two distinct tissues. We can try to choose genes with large sample variance for Y_{ig} but smaller variance for $N_{1,ig}$ and $N_{2,ig}$, and implement a procedure to filter out genes with expression levels that differentiate across tissue components. We split the whole procedure into two stages that respectively

estimates proportions for unknown and referenced components over different gene subsets. A subset of genes with similar expression pattern for N_1 and N_2-components are prepared for degenerating the deconvolution setting from three-component to two-component. In the first stage, the proportions and distribution parameters for the unknown component are estimated. Then we substitute those estimated proportions and parameters into our model, and select a subset of genes that have largest average difference between N_1 and N_2 components for deconvolution. This two-stage approach is arranged as follows.

Stage 1 In this stage, two-component deconvolution is designed for a selected gene subset. We assume a two-component mixture instead of three. N_1 and N_2 components are combined to focus on estimating π_T. We implement it in the following three steps.

• Step 1: A gene subset is selected with small sample standard deviations of reference profiles for both N_1 and N_2-component tissues. Among those genes, we further select genes with $\overline{LN}_{1g} \approx \overline{LN}_{2g}$, where the $\overline{LN}$ is the sample mean for the log2-transformed data.

• Step 2: In the selected gene subset, genes with largest sample standard deviations of Y_g are prepared for the next step.

• Step 3: Run *DeMixT* in two-component setting to estimate μ_{Tg}, σ^2_{Tg} and π_T.

Stage 2 In this stage, $\{\pi_1\}_i$ and $\{\pi_2\}_i$ are estimated in the three-component setting by fixing $\{\pi_T\}_i$ with the estimated values from the first stage.

• Step 1: Genes with most different average reference profiles between N_1 and N_2 components as well as largest sample standard deviations of Y_g are selected.

• Step 2: Run *DeMixT* in three-component setting over the selected genes to estimate π_1 and π_2 given π_T. Finally, given all the estimated parameters, expression

levels $n_{1,ig}$, $n_{2,ig}$ and t_{ig} are reconstituted.

2.3.3 *R*-package

We develop several R functions that are integrated into a freely available R-package *DeMix* (http://bioinformatics.mdanderson.org/main/DeMix:Overview) for *DeMixT* implementation. *DeMixT* contain two functions *DeMixT.S1* and *DeMixT.S2* to run two-step approaches. *DeMixT.S1* estimates tissue-specific proportions for the input expression data, with the probes/transcripts along the rows and samples along the columns. The input expression data of each column is required to be labeled with reference normal profiles or mixed tumor profiles. User can choose if two-stage estimation strategy is employed when a three-component deconvolution is implemented. *DeMixT.S2* estimates the full expression matrices of all consitutents for any input gene subsets with the given tissue-specific proportions for each samples. Finally, the function *DeMixT* combines those two functions to implement the whole pipeline. *DeMixT* can be easily used and integrated to existing pipelines for cancer study.

2.4 Simulations

We include two simulation studies to assess the performance of our algorithm on estimating proportions and distribution parameters respectively for two-component and three-component. In simulation study, we generated 50 replicates of simulation data sets for three-component deconvolution and 40 replicates of simulation for two-component deconvolution.

2.4.1 Simulation design

For testing three-component deconvolution, we simulate expression profiles of 60 samples for each replicate, consisting of 20 N_1-component reference samples, 20 N_2-component reference samples and 20 are mixed tumor samples that needs to be deconvolved.

We generate a combination of proportions for all these mixed samples, $\{\pi_1, \pi_2, \pi_T\}_i$. We assign $\{\pi_1\}_i^S$ with an arithmetic sequence from 0.15 to 0.85 and generate $\{\pi_2\}_i^S$ from an uniform distribution with lower limit to be 0.05 and upper limit to be $0.95 - \pi_{1,i}$ for each samples i. π_T is generated by $1 - \pi_1 - \pi_2$. We generated 200 genes. Expression values of N_1, N_2 and T component tissues for each mixed sample i and gene g are simulated from a Log2-Normal distribution, where $\mu_{N_{1,g}} \sim N(7, 1.5)$, $\mu_{N_{1,g}} \sim N(7, 1.5)$, $\mu_{T,g} \sim N(7, 1.5)$ and $\sigma_{N_{1,g}} = \sigma_{N_{2,g}} = 0.5$, $\sigma_{T,g} = 0.5$ for the first 25 replicates and $\sigma_{T,g} = 1$ for the second 25 replicates. Reference profiles are generated from the same distribution of $N_{1,g}$ and $N_{2,g}$. Then we average simulated expression values according to equation 2.1 using assigned proportions as the weights ,and then generate reference profiles of 20 N_1-component and N_2-component samples. As a result, we create 20 mixed samples along with each 20 reference samples for N_1-component and N_2-component tissues. We repeat this procedure 50 times.

In testing two-component deconvolution, we simulate data in a similar way. $\{\pi_1\}_i^S$ is still generated from an arithmetic sequence from 0.15 to 0.85, then $\pi_{T,i}$ is calcualted from $1 - \pi_{1,i}$. We also generated 200 genes, and $\mu_{N_{1,g}}$ and $\mu_{T,g}$ are generated in the same way but with $\sigma_{N_{1,g}} = \sigma_{T_g} = 0.5$. We simulate expression values of 40 mixed samples with 20 reference samples for N_1-component tissues. We prepare 40 replicates for validation.

For purpose of comparison for three-component deconvolution, we implement a Metropolis−Hastings

algorithm for sampling estimated parameters, which is employed by *DeMixBayes* for two-component deconvolution in the same package but has crazy long running time for three-component deconvolution. Uniform priors are given for $\{\pi_1, \pi_2, \pi_T\}_{i=1}^{S}$. We use the non-informative independent priors assumed in *DeMixBayes* as follows:

$\mu_{N_{1g}} \overset{iid}{\sim} Normal(0, 10,000)$, $\mu_{N_{2g}} \overset{iid}{\sim} Normal(0, 10,000)$, $\mu_{T_g} \overset{iid}{\sim} Normal(0, 10,000)$, $\frac{1}{\sigma^2_{N_{1g}}} \overset{iid}{\sim} Gamma(0.001, 0.001)$, $\frac{1}{\sigma^2_{N_{2g}}} \overset{iid}{\sim} Gamma(0.001, 0.001)$, $\frac{1}{\sigma^2_{T_g}} \overset{iid}{\sim} Gamma(0.001, 0.001)$.

We ran random walk Metropolis-Hastings (RWMH) algorithm for 5000 iterations and thin the chain every 10 iterations to sample $\{\mu_{N_{1g}}, \sigma_{N_{1g}}\}_{g=1}^{G}$, $\{\mu_{N_{2g}}, \sigma_{N_{2g}}\}_{g=1}^{G}$, $\{\mu_{T_g}, \sigma_{T_g}\}_{g=1}^{G}$ from their full conditional posterior and calculate their posterior means.

2.4.2 Performance evaluation

We evaluate the performance of our method by running *DeMixT* without two-stage optimization, that finishes very quickly in two-component setting. For both two-component and three-component deconvolution, our method has small biases and high correlations with truth. Esimates for the same true proportion from different replicates are highly stable. (Fig. 2.3)

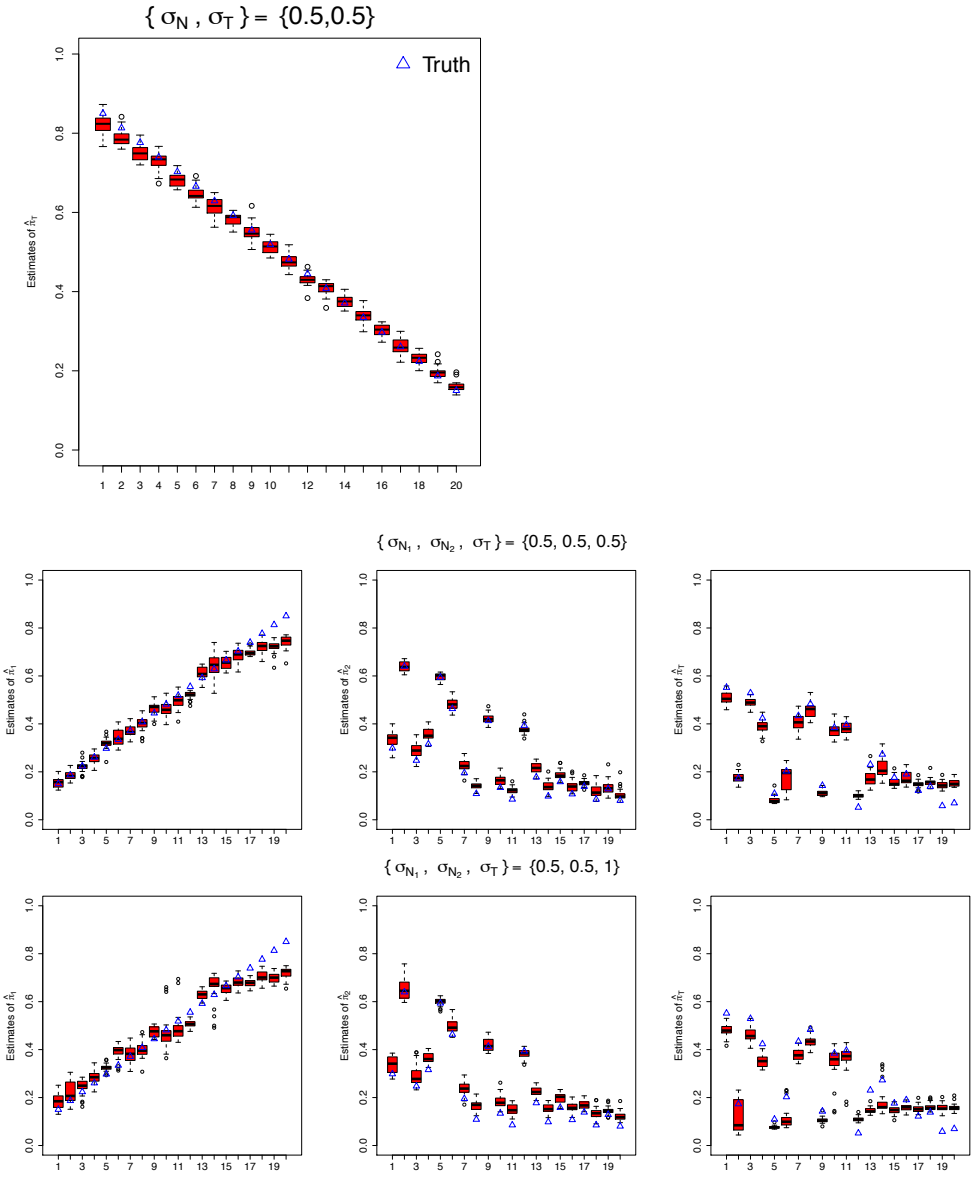


Figure 2.3 : Boxplots of estimated proportions from *DeMixT* for 20 samples in all the simulation replicates. Blue triangles are the truth. The top plot gives π_1 estimates for two-component deconvolution; the bottom plot gives estimates of π_1, π_2 and π_T for three-component deconvolution.

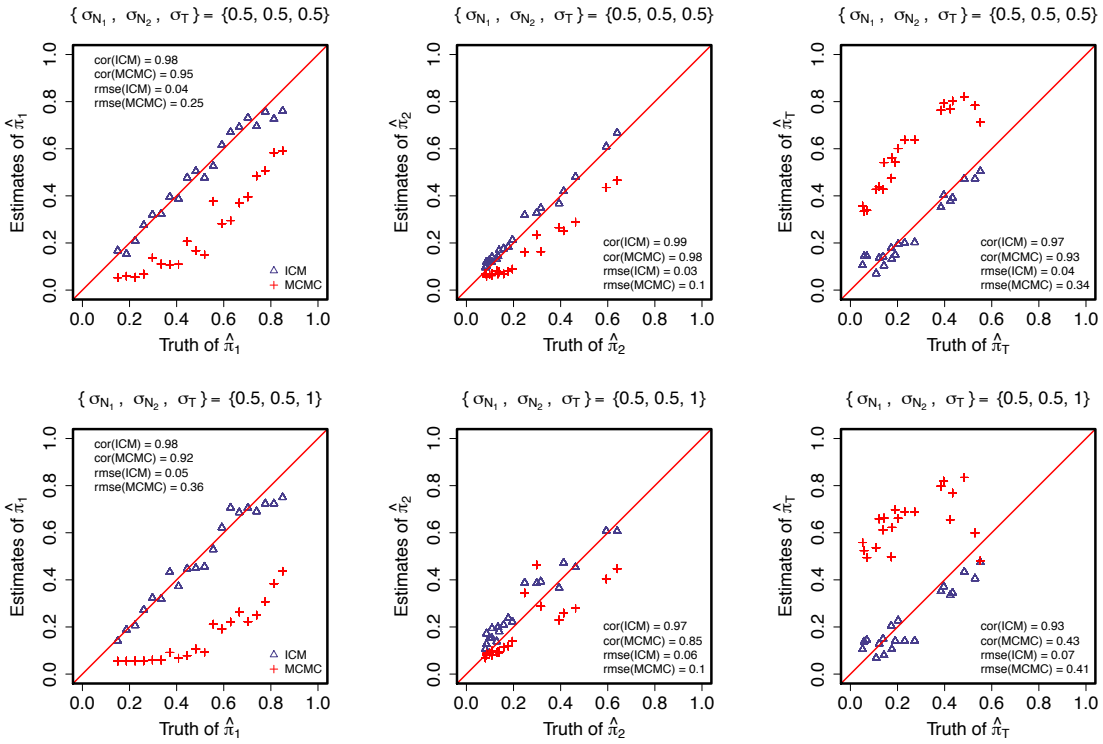


Figure 2.4 : Estimated $\{\pi_1, \pi_2, \pi_T\}$ versus truth in three-component deconvolution simulations through *DeMixT* and RWMH; red dots denote proportion estimates from RWMH; blue dots denote proportion estimates from *DeMixT*.

We compare proportion estimates between those two different methods for three-component deconvolution. We compute the Pearson correlation coefficients (COR) and root-mean-square error (RMSE) between estimates and truth. Our results give COR above 0.96 and RMSE below 0.05 in the first 25 replicates when $\sigma_{T,g} = 0.5$. In the second 25 replicates with larger $\sigma_{T,g}$, our method still gives a substantially better COR and smaller RMSE with truth than RWMH (Fig. 2.3 and 2.4). Furthermore the numerical integration undergirding the two-component model in *DeMixBayes* is too computationally intensive to be feasible for the three-component model. The method of iterated conditional modes can quickly converge and cost no more than 10 hours, while the Bayesian methods require more than one whole week for running

5000 iterations.

2.5 Experiment Validation

2.5.1 Measures for evaluation

For evaluating the performance of our experiment validation, we first provide and define several measures we will use in the following discussion.

Concordance correlation coefficient

In our expriment validation, we use the concordance correlation coefficient (COR) and root-mean-squared error (RMSE) to evaluate the performance of our method (Lawrence and Lin, 1989). The COR ρ_{xy} is a measure of the agreement between two variables x and y and is defined as $\rho_{xy} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2}$, where μ and σ^2 are the corresponding mean and variance for each variable.

Summary statistics for measure of reproducibility

To assess the stability of the performance across different scenarios/schemes for different methods, we calculated the sample standard deviations of the estimation error $\hat{\pi}_i - \pi_i$ across different scenarios for each observation and then averaged over them for all the samples. We define a reproducibility statistics as $R = \frac{1}{N}\sum_n^N(\frac{1}{T-1}\sum_t^T(\epsilon_n^t - \frac{1}{T}\sum_t^T \epsilon_n^t)^2)^{\frac{1}{2}}$, where $\epsilon_n^t = \hat{\pi}_n^t - \pi_n$. $\hat{\pi}_n^t$ is the estimated value for the n-th observation and t-th scenario and π_n is the truth. N denotes the sample size and T is the number

of scenarios. It measures the variations of estimation errors across different scenarios, so a method with smaller R has more reproducibility in estimation across different scenarios.

A deconvolution tool for comparison: *ISOpure*

We introduce another published computational deconvolution tool, ISOpure, which can also jointly estimate tissue-specific proportions and tumor expression profiles for biospecimens mixed of more than two component tissues (Quon et al., 2013). This method is similar to *DeMixT* in assuming a cancer component that does not require reference profiles. Actually *ISOpure* and *DeMixT* are these several only methods that are able to deconvolve tumor profiles without requiring reference profiles from all the mixing components. The model assumption in *ISOpure* is provided as follows:

$$t_n = a_n c_n + (1 - a_n) h_n + e_n \tag{2.7}$$

t_n is tumor profile, c_n is the component cancer profile, e_n is the reconstruction error, a_n is the fraction of tumors, and $h_n = \sum_{r=1}^{R} \theta_{n,r} b_r$, which is a weighted average of available healthy tissue profiles b_r.

Our method outperforms *ISOpure* in three aspects: First, *ISOpure* assumes a convex combination of reference healthy tissue profiles for the normal profiles. So it does not explicitly model sample variations for normal profiles, a feature that is necessary for estimating individual normal profiles. Second, *DeMixT* provides more reasonable variation estimates across genes, whereas ISOpure seems to underestimate the gene-specific variances. Third, in our next discussion, we will show that *DeMixT* is able to

provide more accurate estimates of tissue-specific proportions and mean expression values.

ISOpure has been wrapped into a *MATLAB* and an *R* package, so it is convenient to use it to compare with our method. In our following exploration, we will use *ISOpure* as a tool for comparison.

2.5.2 Microarray data analysis of mixed two-component tissues

We first validate our esimation performance for tissue proporitons in the two-component deconvolution setting. We downloaded two microarray datasets with GSE5350 from GEO browser. These dat sets mixed RNA sample from isolated 100% Stratagene Universal Human Reference RNA (A) and 100% Ambion Human Brain Reference RNA (B) at 75% : 25% ratio and 25% : 75% ratio. Ten mixed samples and five reference samples processed from two test sites are respectively prepared for deconvolution, and they are denoted as MAQC1 and MAQC3 (Shi et al., 2006).

We first selected probes with low background noise following a previous procedure. After running our package on those samples, we compared our proportion estimates versus truth. The scatter plots show that we can estimate the proportions for all the samples very well with high COR and low RMSE for both MAQC1 and MAQC3 data (Fig. 2.5).

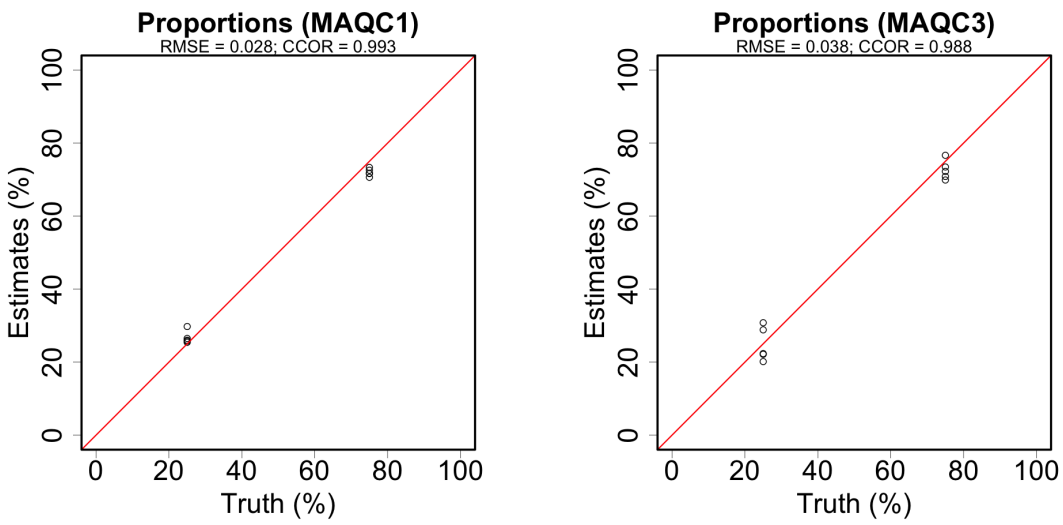


Figure 2.5 : Esimation of proportions of the unknown component for MAQC1 data and MAQC3 data. Estimated tissue proportions are compared versus true proportions.

2.5.3 Microarray data analysis of mixed three-component tissues

We downloaded datasets GEO19830 (Shen-Orr et al., 2010) from the GEO browser as benchmark data

($http://www.ncbi.nlm.nih.gov/geo/browse$) to evaluate the performance of $DeMixT$ on microarray data. This microarray experiment was designed for expression profiling of samples from Rattus norvegicus with the Affymetrix Rat Genome 230 2.0 Array. The data sets include 30 mixed samples of liver, brain and lung tissues in 10 different mixture proportions (Table A.1). Three technical replicates were prepared for each mixture proportion and for the pure liver, brain and lung tissues with purity of 100%. We downloaded the CEL files and used the R package $\{affy\}$ to summarize the raw probe intensities with quantile normalization but without background correction according to robust multi-array average (RMA) procedures because background cor-

rection is thought to possibly alter the linearity (Liebner et al., 2014; Bolstad et al., 2003; Irizarry et al., 2003; Carvalho et al., 2007).

To confirm the linearity assumptions for probes that are input into *DeMixT*, we required probes prepared for our deconvolution to be measured in the upper quartile of the log2-transformation data with both N_1-component and N_2-component tissues. The linear relationship fits well for our selected probes (green dots) in the experimental validation of the microarray data, from which the log2-transformed expression level of at least two tissues measured above 7 for the average level (Fig. A.1). We filtered out probes with low intensity after checking for the linearity assumption; this was likely caused by technical measurement errors and background noise.

Linearity testing

Our model relies on an assumption that the tissue-specific expression levels are mixed linearly. We can check for the validity of this assumption when $\{\pi_1, \pi_2\}_i$ are known. By making simple transformations of equation 2.1, we have

$$Y_{ig} = \pi_{1,i}N_{1,ig} + \pi_{2,i}N_{2,ig} + (1 - \pi_{1,i} - \pi_{2,i})T_{ig} \Leftrightarrow \begin{cases} \pi_{1,i} = \frac{Y_{ig} - T_{ig} - \pi_{2,i}(N_{2,ig} - T_{ig})}{N_{1,ig} - T_{ig}} \\ \pi_{2,i} = \frac{Y_{ig} - T_{ig} - \pi_{1,i}(N_{1,ig} - T_{ig})}{N_{2,ig} - T_{ig}} \end{cases} \quad (2.8)$$

Thus, we created scatter plots with a regression line to compare $Y_{ig} - T_{ig} - \pi_{2,i}(\bar{N}_{2,g} - T_{ig})$ with $\bar{N}_{1,g} - T_{ig}$ and $Y_{ig} - T_{ig} - \pi_{1,i}(\bar{N}_{1,g} - T_{ig})$ with $\bar{N}_{2,g} - T_{ig}$, where the sample mean for $\bar{N}_{1,g}$(e.g. Liver), $\bar{N}_{2,g}$(e.g. Brain) and $\bar{T}_g$(e.g. Lung) were used instead of each $N_{1,ig}$, $N_{2,ig}$ and T_{ig}.

Estimation of tissue proportions

The samples in the data set GSE19830 are all mixtures of pure liver, brain and lung tissues in specific ratios. We used these samples with 100% purity as pure N_1-component, N_2-component and T-component tissue samples. The deconvolution for the remaining 30 mixed samples was performed in three different schemes, respectively assuming liver, brain and lung tissues to be the unknown T-component tissue. We implemented *DeMixT* on those mixed samples and evaluated the estimation performance for tissue proportions and expression levels. In the first stage of the two-stage estimation, we selected 250 probes to be deconvolved on *DeMixT* after running 10 iterations, and in the second stage, we selected 200 probes. For comparison purposes, we ran *ISOpure* to estimate the tissue proportion for the T-component tissue.

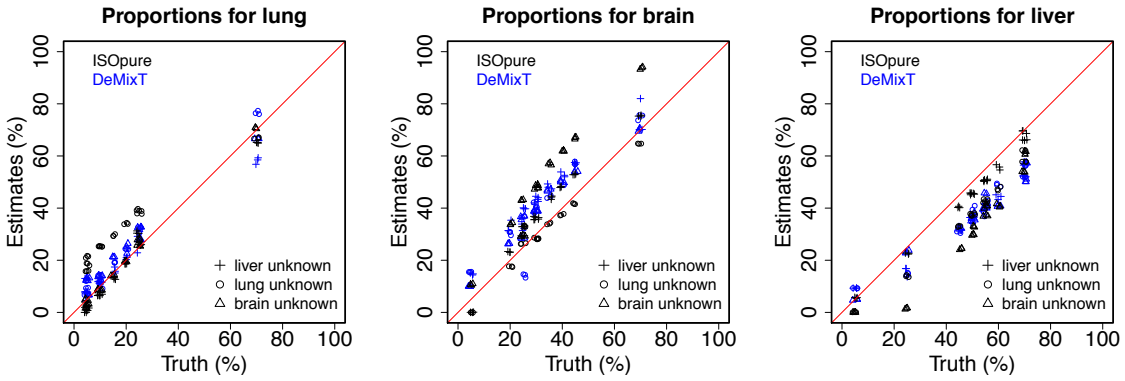


Figure 2.6 : Consistency in estimation of rat tissue proportions. Scatter plots of estimated tissue proportions against true tissue proportions when either the liver, brain, or lung tissue is assumed to be the unknown tissue; blue rectangles represent*DeMixT* estimates when liver tissue is assumed to be unknown; blue circles represent*DeMixT* estimates when lung tissue is assumed to be unknown; blue rectangles represent *DeMixT* estimates when brain tissue is assumed to be unknown; black crosses represent *ISOpure* estimates.

For all those three deconvolution schemes aforementioned, *DeMixT* present high

correlations between estimates and truth for $\{\pi_1, \pi_2, \pi_T\}_{i=1}^{S}$ and RMSE of estimates across these three scenarios were all reported less than 0.15 (Table 2.2 and Table 2.3). In Table 2.3 and Figure 2.6, *DeMixT* was reported with much smaller RMSE below 0.1 for the unknown T-component than *ISOpure* when brain and lung are assumed to be the T-component tissues. While *DeMixT* was observed with larger RMSE when liver is assumed to be unknown (Fig. 2.6), the performance across all the three scenarios are consistent where liver was reported with small under-estimation, brain with small over-estimation and lung with accurate estimation (Table 2.4).

Table 2.2 : Concordance correlation coefficients between estimated proportions and true proportions in the GSE19830 data set. The 95% confidence interval is given in the bracket.

Estimated Tissue	Brain	Lung	Liver	Average
DeMixT (Brain Unknown)	0.88 (0.80, 0.93)	0.95 (0.91, 0.97)	0.74 (0.61, 0.83)	0.86
DeMixT (Lung Unknown)	0.84 (0.71, 0.91)	0.97 (0.95, 0.98)	0.75 (0.63, 0.84)	0.85
DeMixT (Liver Unknown)	0.77 (0.65, 0.86)	0.96 (0.94, 0.97)	0.74 (0.62, 0.83)	0.82
ISOpure (Brain Unknown)	0.69 (0.55, 0.79)	1 (1.00, 1.00)	0.72 (0.58, 0.81)	0.8
ISOpure (Lung Unknown)	0.97 (0.94, 0.99)	0.74 (0.61, 0.83)	0.84 (0.75, 0.90)	0.85
ISOpure (Liver Unknown)	0.93 (0.88, 0.96)	0.98 (0.96, 0.99)	0.98 (0.96, 0.99)	0.96

Table 2.3 : Root mean squared errors (RMSEs) between estimated proportions and true proportions in the GSE19830 data set.

Estimated Tissue	Brain	Lung	Liver	Average
DeMixT (Brain Unknown)	0.08	0.06	0.13	0.09
DeMixT (Lung Unknown)	0.1	0.05	0.13	0.09
DeMixT (Liver Unknown)	0.12	0.05	0.13	0.1
ISOpure (Brain Unknown)	0.18	0.02	0.16	0.12
ISOpure (Lung Unknown)	0.04	0.14	0.11	0.1
ISOpure (Liver Unknown)	0.07	0.04	0.04	0.05

Estimation of tissue-specific expression

Corresponding to each scheme we performed deconvolution, we also estimated tissue-specific expression levels for the whole probe set through *DeMixT* by fixing $\{\pi_1, \pi_2, \pi_T\}_i$ with the estimates of tissue proportions. Expression levels for each probe were estimated independently when the tissue proportions were given. Both *DeMixT* and *ISOpure* yielded accurate estimations of the mean expression values when deconvolving the gene expression values (Fig. 2.7).

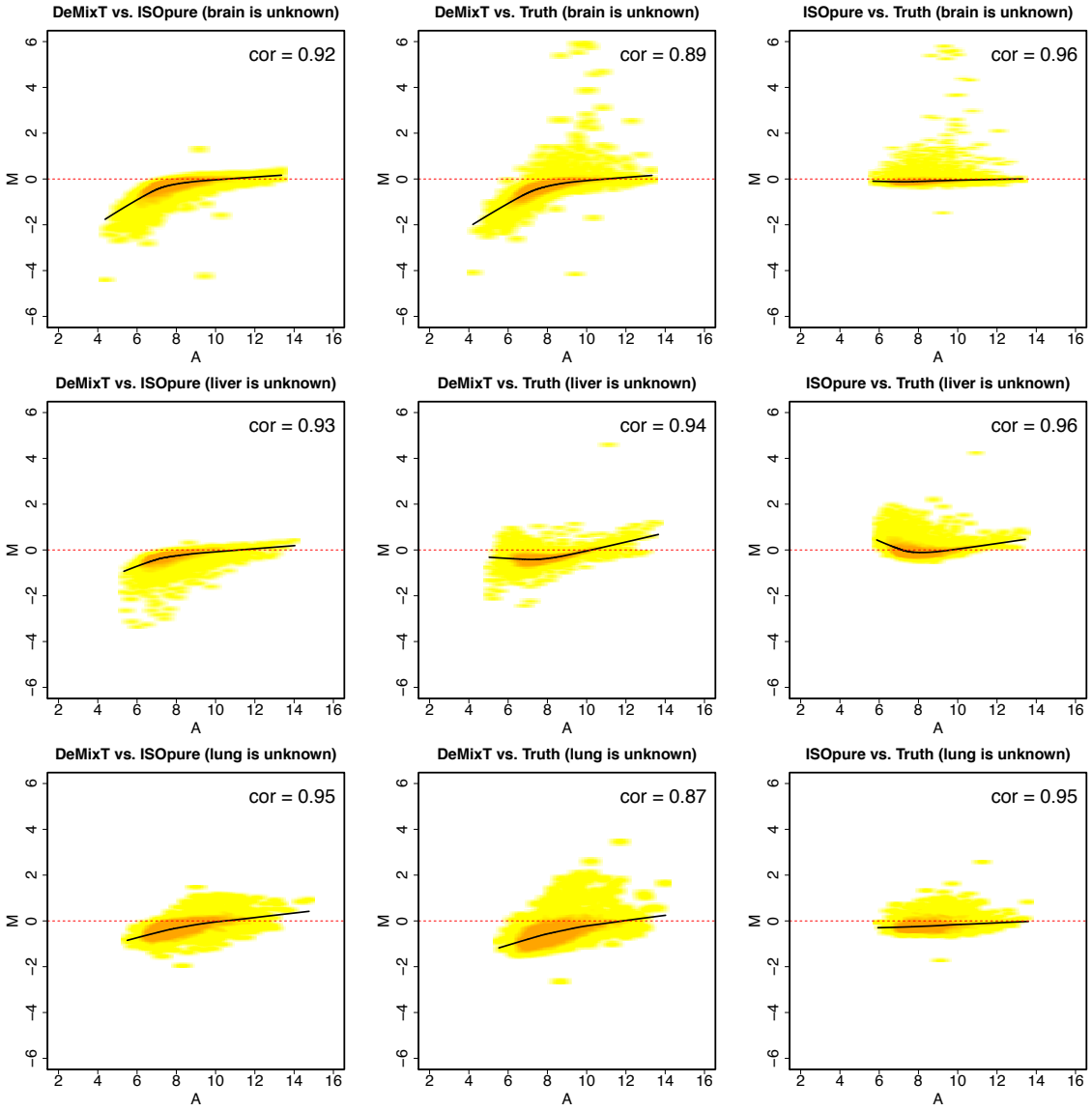


Figure 2.7 : MA plots of estimated tissue-specific expression between *DeMixT* and *ISOpure* in the GSE19830 data set. *DeMixT* provides accurate estimation of tissue-specific expression. MA plots compare the mean values of deconvolved expression levels across genes for*DeMixT* vs. *ISOpure*, *DeMixT* vs. observed samples, and *ISOpure* vs. observed samples when either liver, lung or brain tissue is assumed to be the unknown component.

Table 2.4 : Calculated value of summary statistics of reproducibility for estimation of component proportions across different scenarios in the GSE19830 data set and RNA-seq data from mixed cell line experiment. H1092: lung tumor adenocarcinoma; CAF: cancer-associated fibroblasts; TIL: tumor infiltrating lymphocytes.

Estimated Tissue	*DeMixT*	ISOpure
Brain	0.03	0.10
Lung	0.03	0.08
Liver	0.03	0.07
H1092	0.03	0.39
CAF	0.02	0.40
TIL	0.02	0.20

2.5.4 RNA-seq mixed cell line experiment

The experiment is designed to validate the performance of *DeMixT* on RNA-seq data sets. The expression scale of raw level microarray is similar to that of sequencing data. Although our log2-normal distribution assumption is sourced from microarray data, the concordance between sequencing data and microarray data on transcript abundance suggests that it can be directly applied to RNA-seq data (Wang et al., 2014). We mixed mRNA from cell lines of lung adenocarcinoma in humans (H1092), cancer-associated fibroblasts (CAFs) and tumor infiltrating lymphocytes (TIL) in different cell proportions to generate 23 samples with two replicates for each sample (Table A.2). The RNA amount of each tissue in the mixture samples was calculated on the basis of real RNA concentrations tested in our collaborators' lab. We prepared three pure cell line samples with cell proportions of 100% for lung tumor, pure fibroblasts, and pure lymphocytes. Each sample was generated with two replicates.

The following pipeline was devised to obtain gene expression counts for these samples. At the onset, raw reads generated from pair-ended sequencing were mapped to the genome reference. The mapped reads were cleaned and sorted by their names to give the reads random ordering. For short-read alignment, we applied the *R* packages *GenomicFeatures* and *GenomicRanges* and used the reference human genome *h19* to generate a reference table. We used the function *findOverlaps* to obtain gene expression counts for all the samples. Then we input the gene expression counts of the mixed samples and pure samples from every two components of tissue to validate the deconvolution.

Estimation of tissue proportions

We treated every replicate of a sample as an individual sample, so we had 46 mixed tumor samples to deconvolve. Each type of mixed tissue was assumed to be the unknown $T-$type tissue. Before deconvolution, we scale-normalized the data matrix and discarded the genes that contained a count of zero. In both stages of our optimization, we selected 250 genes for deconvolution in the respective two-component and three-component settings. We evaluated the performance of *DeMixT* on RNA-seq data the same way by comparing it with the true mRNA proportions and estimates from *ISOpure*.

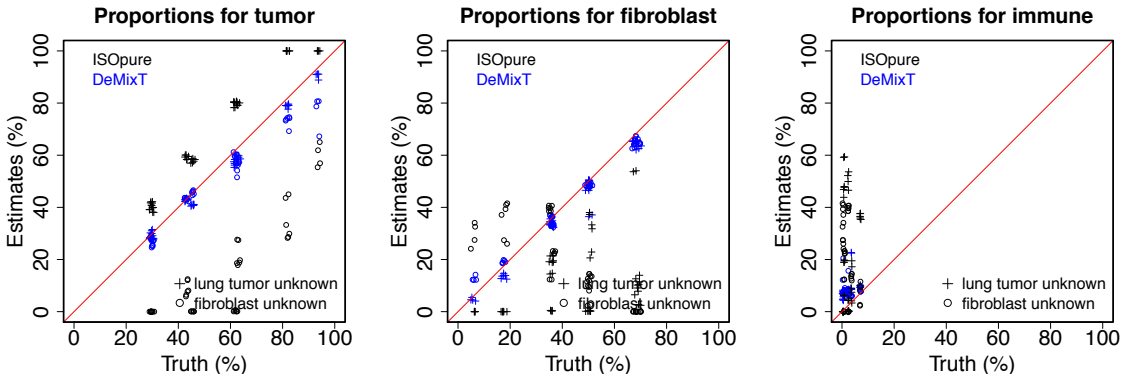


Figure 2.8 : *DeMixT* yields accurate estimation of proportions of RNAseq data generated from mixed lung cancer cell lines. Scatter plots of estimated tissue proportion against true tissue proportion when either lung tumor or fibroblast is assumed to be the unknown T-type tissue; blue crosses represent *DeMixT* estimates when lung tumor cell is assumed to be unknown; blue rectangles represent *DeMixT* estimates when fibroblast cell is assumed to be unknown; black crosses and rectangles represent *ISOpure* estimates.

When H1092 and CAFs are considered as the unknown component, *DeMixT* provides expected estimates with tighter COR and remarkably smaller RMSE (< 0.1) with the true proportions compared with *ISOpure* (Table 2.5 and Table 2.6). Proportion estimates given by *DeMixT* are consistent when different components were treated as unknown in our experiments (Fig. 2.8). In our experiment, the truth mRNA proportions of TIL are all very small, so it is too difficult to detect such weak signals without reference profiles for TIL. When TIL is considered as the T-component, *DeMixT* and *ISOpure* are both reported with poor performance by overestimating the proportions of immune tissues (Fig. 2.9). When the proportions of one component tissue are remarkably low($< 5\%$) for all the samples, mixed samples are artifacts of three-component mixtures without any reference information for the low proportion component, because this component could be recognized as noises. Deconvolution algorithm will fake its information for the unknown component, ex-

cept that it has much higher or lower expression values for all the selected genes than that of other components.

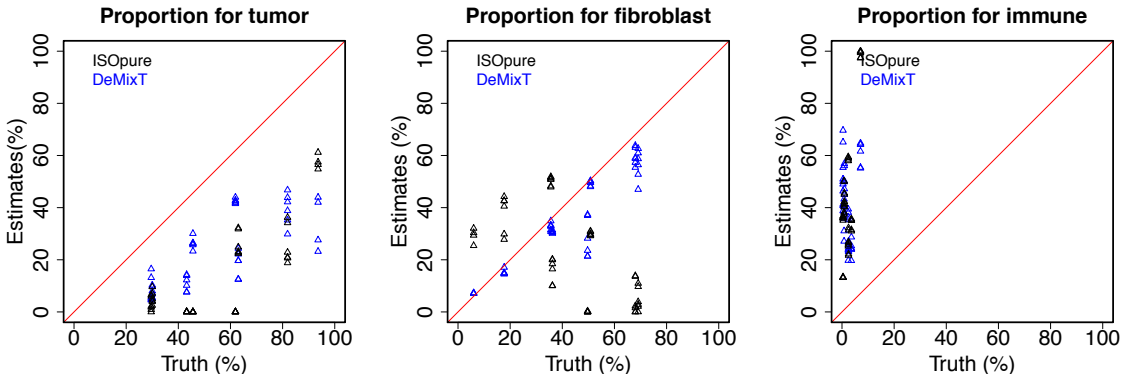


Figure 2.9 : Estimated Tissue proportion versus truth proportion when cell type H1092 is the unknown T-component tissue; Red dots represent *DeMixT* estimates; Blue dots represent *ISOpure* estimates.

Table 2.5 : Concordance correlation coefficients between estimated proportions and true proportions in RNA-seq data from mixed cell line experiment. The 95% confidence interval is given in the bracket. We use H1092, CAF and TIL to respectively denote lung tumor adenocarcinoma, cancer-associated fibroblasts and tumor infiltrating lymphocytes.

Estimated Tissue	Lung Tumor (H1092)	Fibroblast (CAF)	Immune (TIL)	Average
DeMixT (H1092 Unknown)	0.99 (0.98, 0.99)	0.98 (0.96, 0.99)	0.12 (0.03, 0.20)	0.7
DeMixT (CAF Unknown)	0.94 (0.90, 0.96)	0.99 (0.98, 0.99)	-0.02 (-0.10, 0.06)	0.64
ISOpure (H1092 Unknown)	0.81 (0.73, 0.87)	0.14 (0.03, 0.25)	0.03 (0, 0.07)	0.33
ISOpure (CAF Unknown)	0.28 (0.18, 0.37)	0.62 (0.51, 0.71)	-0.04 (-0.09, 0)	0.29

Table 2.6 : Root mean squared errors between estimated proportions and true proportions in RNA-seq data from mixed cell line experiment.

Estimated Tissue	Lung Tumor (H1092)	Fibroblast (CAF)	Immune (TIL)	Average
DeMixT (H1092 Unknown)	0.03	0.04	0.07	0.05
DeMixT (CAF Unknown)	0.07	0.03	0.06	0.05
ISOpure (H1092 Unknown)	0.14	0.37	0.26	0.26
ISOpure (CAF Unknown)	0.42	0.26	0.22	0.3

Estimation of tissue-specific expression

For these two schemes we are able to accurately estimate proportions, we estimated tissue-specific expressions for all the genes without zero count by substituting estimates of $\{\pi_1, \pi_2, \pi_T\}_{i=1}^{S}$. We evaluate deconvolved expression profiles in the log2-transformed scale.

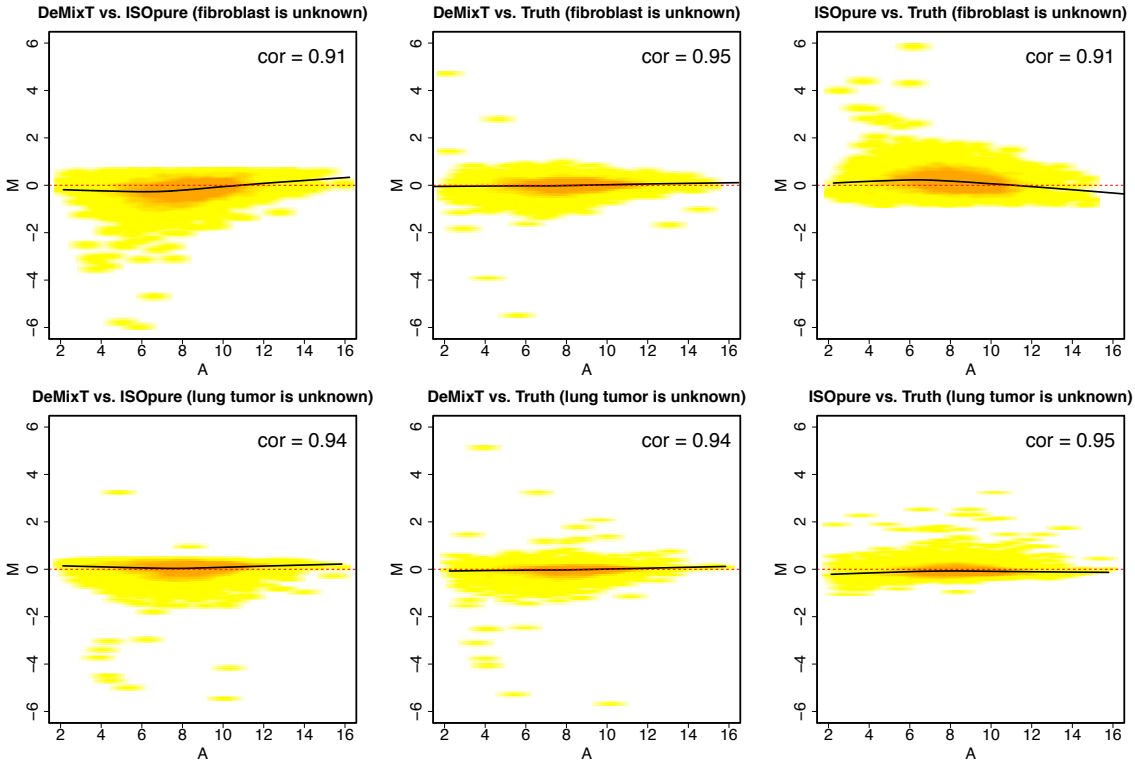


Figure 2.10 : MA plots of estimated tissue-specific expression between *DeMixT* and *ISOpure* in mixed cell line RNA-seq data. Improperly estimated probes from *DeMixT* are removed. *DeMixT* provides accurate estimation of tissue-specific expression. MA plots compare mean value of deconvolved expression profiles for *DeMixT* vs. *ISOpure*, *DeMixT* vs. observed samples, and *ISOpure* vs. observed samples when either lung tumor or fibroblast cell is assumed to be the unknown component.

It shows high correlation between mean values of deconvolved expression profiles and the measured true mean values (Fig. 2.10). Much lower RMSE demonstrates better performance of *DeMixT* for deconvolving sequencing data, and it validates the use of our model on sequencing data.

2.6 Real Data Analysis

2.6.1 Microarray study of laser capture microdissected prostate cancer patient samples

We applied *DeMixT* to real tumor samples and examined its performance. Laser-capture microdissection was used to separate stromal and tumor tissues from prostate cancer samples, providing a gold-standard validation data set. We collected a cohort of samples from prostate cancer patients, which consist of 25 samples of isolated tumor tissues, 25 samples of isolated stromal tissues and 23 mixture samples. Radical prostatectomy specimens were annotated in detail by pathologists, and regions of interest were identified that corresponded to the benign epithelium, PIN and tumor, each with its surrounding stroma. These regions were laser-capture microdissected using ArcturusXT system (Life Technologies). Additional areas of admixed tumor and adjacent stromal tissue were taken. RNA was extracted by AllPrep (Qiagen) and quantified by RiboGreen assay (Life Technology). RNA labeling was performed using SensationPlus FFPE method (Affymetrix) and hybridized to Affymetrix Gene Array STA 1.0. For the analysis, we used a subset of RMA normalized gene expression data corresponding to tumor, tumor-adjacent stromal tissue and an admixed region. We tested deconvolution in these settings when the tumor is unknown and the stromal tissue is unknown. An explicit calculation of the average expression difference between the two components $|\overline{N} - \overline{T}|$, where $\overline{N}$ and $\overline{T}$ are sample means of normal and tumor tissues, shows there is just a small portion of probes with differential expression. To validate our deconvolution method, we pre-selected a subset of the top 80 differentially expressed probes with the largest $|\overline{N} - \overline{T}|$ and used this pre-selected probe set for deconvolution.

Result analysis

After preselecting a subset of probes for deconvolution, we ran an analysis with either tumor or stromal components treated as unknown. We found that *DeMixT* obtained concordant estimates of tumor purity under the two conditions ($r = 0.87$) while *ISOpure* did not ($r = 0.36$) (Fig. 2.11). *DeMixT* also tended to provide mean component-specific expression levels with much lower biases than ISOpure (Fig. 2.12) and yielded standard deviation estimates that were close to those from the dissected tumor samples (Fig. 2.13).

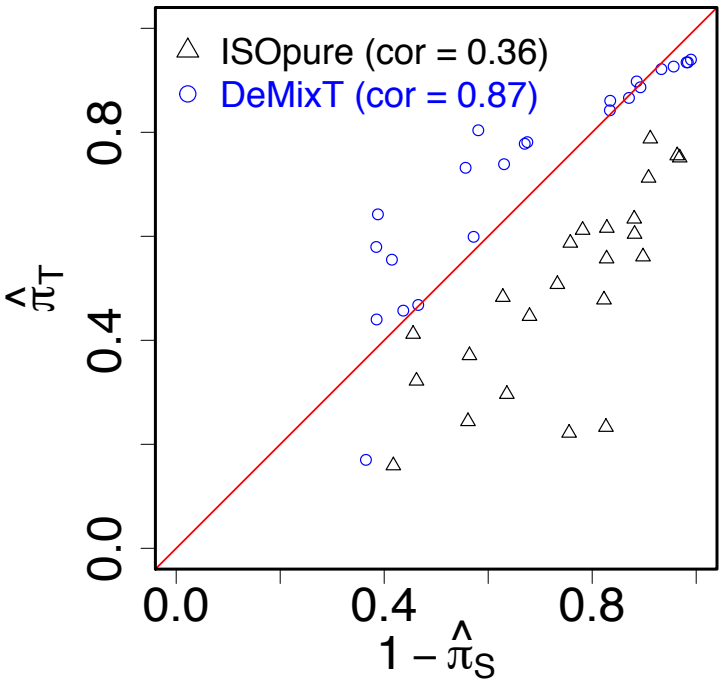


Figure 2.11 : Scatter plot of estimated tumor proportion when the tumor proportion is unknown against those when the stromal proportion is unknown in prostate cancer patient samples; estimations of*DeMixT* (blue) are compared with those of *ISOpure* (black).

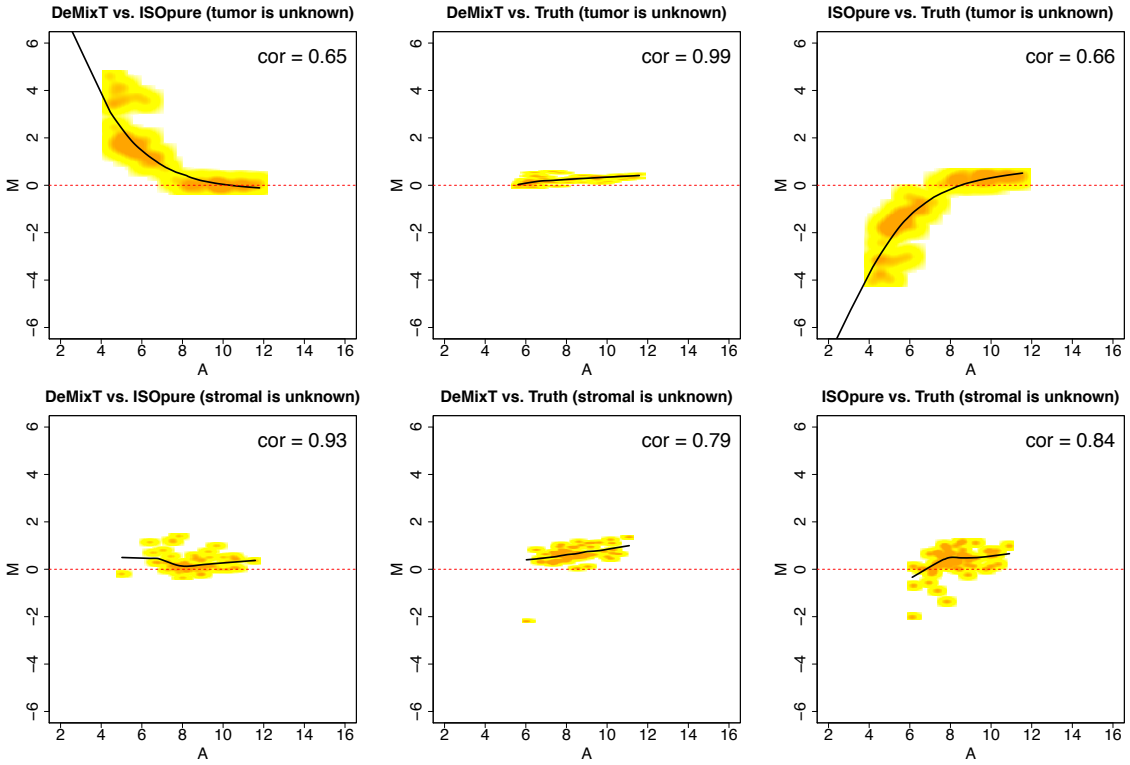


Figure 2.12 : MA plots of estimated tissue-specific expression between *DeMixT* and *ISOpure* in a microarray study of prostate cancer samples. MA plots compare mean value of deconvolved expression profiles for *DeMixT* vs. *ISOpure*, *DeMixT* vs. observed samples, and *ISOpure* vs. observed samples when either tumor or stromal tissue is assumed to be the unknown component. We used a filtered probe subset with the most differential expression between tumor and stromal tissues and smaller expression variation for known tissues from 23 lung cancer patient samples.

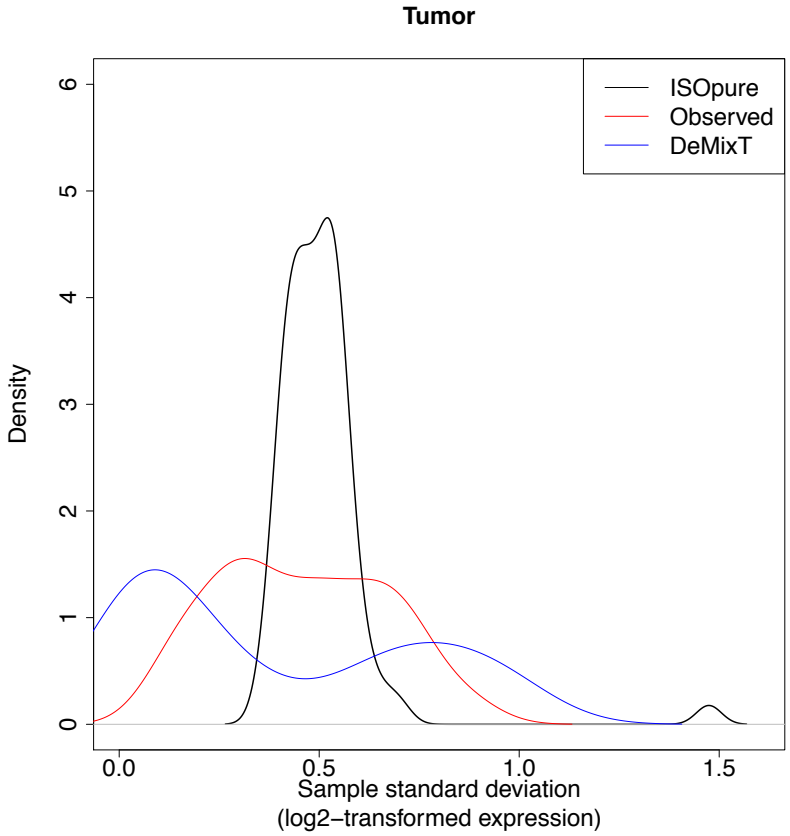


Figure 2.13 : Density plot comparing sample standard deviations between deconvolved expression profiles of subset probes for *DeMixT* and *ISOpure* when tumor tissue is assumed to be the unknown component; with measured expression profiles of isolated tumor tissues.

2.6.2 Immune infiltration in virus-associated tumors

Infection with the human papillomavirus (HPV) plays a critical role in cervical cancer and likely in a subset of head and neck squamous cell carcinoma (HNSC). HNSC tumors with HPV+ test results are believed to be clinically distinct from tumors with HPV- test results. A recent study has demonstrated that the infiltration of immune cells, both lymphocytes and myelocytes, is positively associated with viral infection in virus-associated tumors, where the high viral infection group corresponds

with high rates of immune infiltration (Li et al., 2016). To validate this finding, we downloaded HNSC RNA-seq data from the TCGA data portal (Network et al., 2015a) and ran *DeMixT* to estimate the proportions of tumor cells, stromal cells and immune cells.

Data analysis

We devised an estimation algorithm for estimating the immune cell proportions (Fig. 2.14) and downloaded RNA-seq data for head and neck squamous cell carcinoma (HNSCC) from The Cancer Genome Atlas (TCGA) data portal ($https://tcga-data.nci.nih.gov/tcga/$). We also collected the infection information of human papillomavirus (HPV) for HNSCC samples, so that we have the positive infected group and the negative infected group for each cancer. After removing mislabeled samples, we have 287 samples (44 normal, 243 tumor) for HNSCC. We scale normalized the expression data for each cancer type and filtered out genes with zero count in any sample.

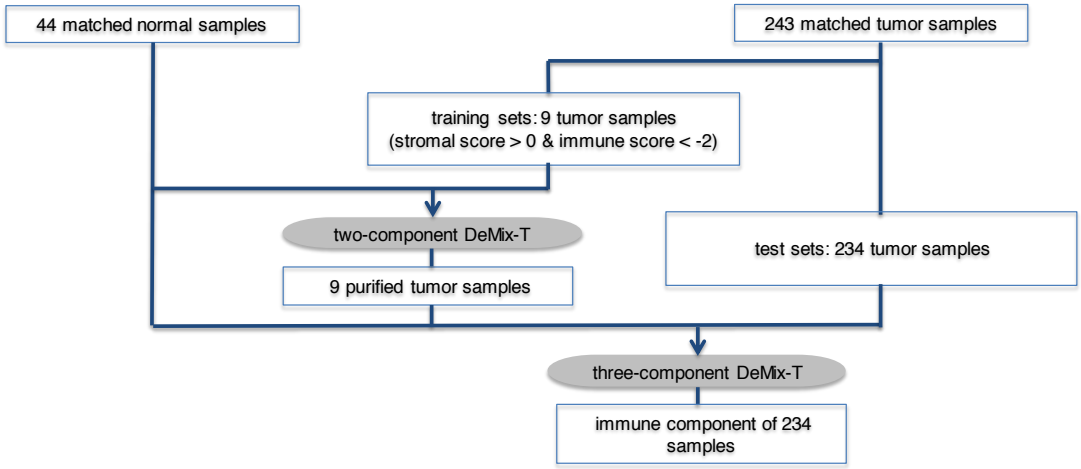


Figure 2.14 : Data analysis workflow for validation of immune infiltration in HNSC tumors. We assigned 43 tumor samples in the training set and 230 tumor samples in the test set. We use *DeMixT* in a two-component setting and a three-component setting in different steps.

We designed a pipeline to prepare the expression profiles of pure tumor tissues for deconvolution. That process can be separated into two steps and we filtered out different gene subsets to run two-component or three-component*DeMixT* in each step. We downloaded stromal and immune scores from single-sample gene set enrichment analysis (ssGSEA) for all the TCGA samples (Network et al., 2015a) and used them to help us group the samples. We pre-selected 9 tumor samples with positive stromal scores (> 0) and negative immune scores (< -2), which implies enriched stromal tissues with low immune infiltration. We considered these samples as a mixture of

just stromal and tumor tissues and ran the two-component deconvolution algorithm in *DeMixT* to obtain deconvolved expressions of pure tumor tissues for each individual sample. deconvolved expression profiles from the genes that have small variations were combined as the input for deconvolution of all the remaining tumor samples. In the next step, all the remaining tumor samples were deconvolved using expression matrices of normal samples and reconstituted tumor component from the first step. With estimated tumor, stromal and immune proportions for each cancer type, we made several analyses by comparing them between the positive and negative virus infection groups.

Result analysis

We compared the estimated immune proportions in the test set between the tumor groups, one with positive and one with negative HPV infection results. The boxplot (Fig. 2.15) and the density plot (Fig. 2.16) show that tumor samples with HPV+ test results had systematically higher immune component proportion estimates than those with HPV- test results ($P = 0.004$). We also analyzed the deconvolved expressions of three important immune cell-related genes, CD4, CD8A and HLA-DQB1 that are expressed in the immune cells. This result (Fig. 2.17) shows that deconvolved expression levels are higher in the immune component than in the other two components for those three genes.

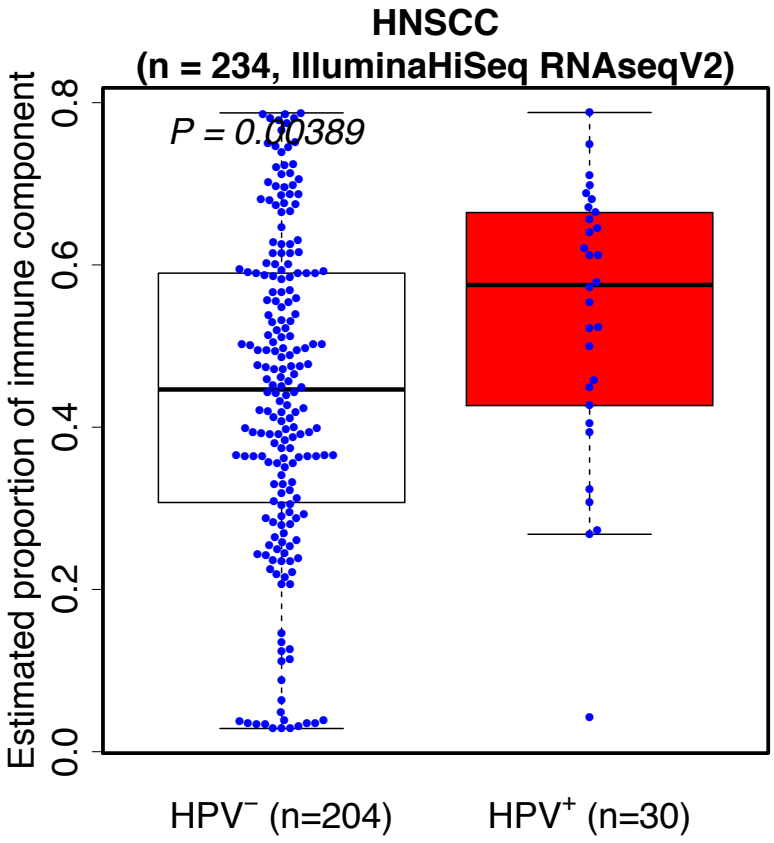


Figure 2.15 : Box and whisker plots of immune proportions HNSC samples in the test set display differences between HPV^+ (red) and HPV^- (white) samples

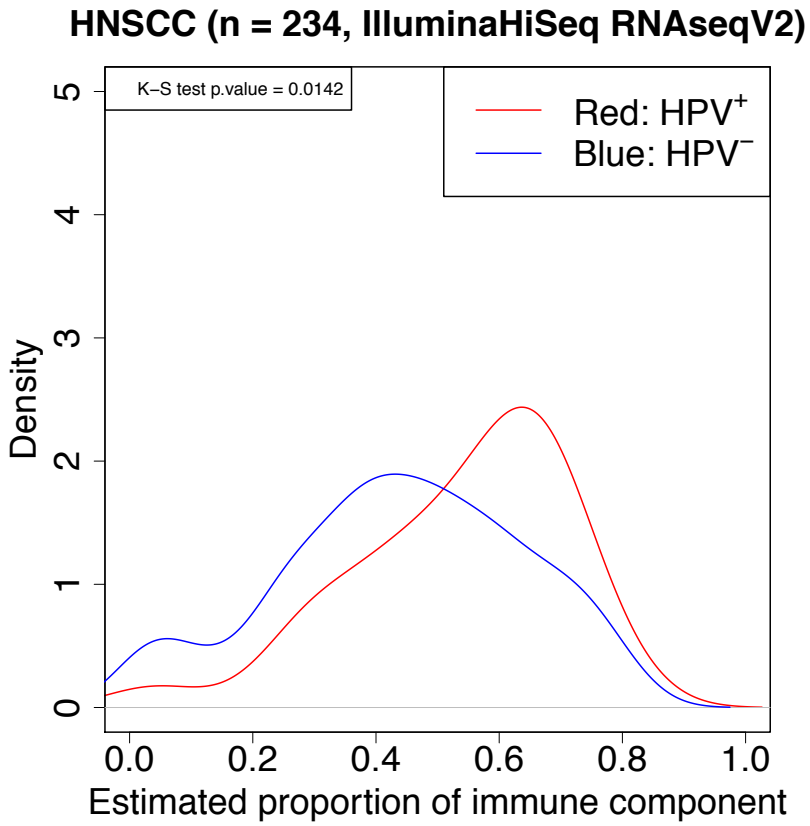


Figure 2.16 : Density plot of immune proportions for tumor samples in test set with HPV test: Red line is for estimated immune proportions of tumor samples with positive HPV test; blue is for those with negative HPV test. From the probability density plot, we observe that the tumor samples with HPV-positive test results have more mass in the region of high immune proportion than those with HPV-negative test results.

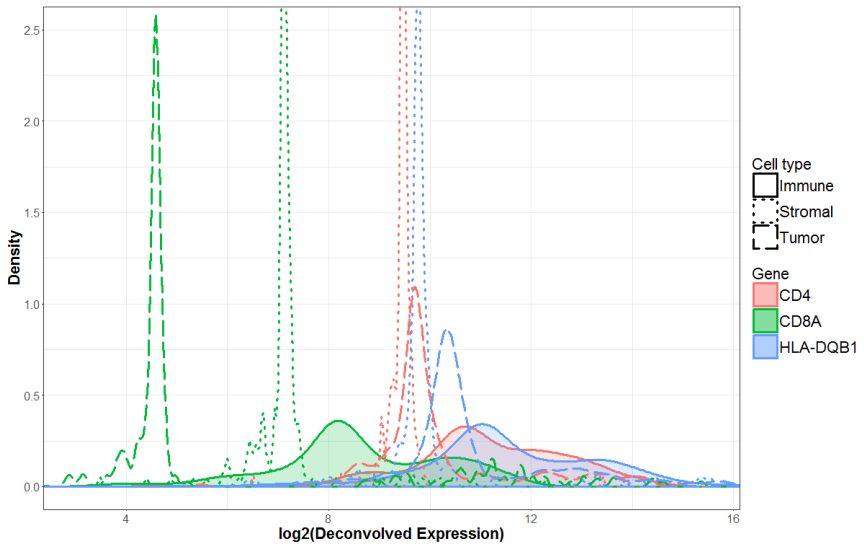


Figure 2.17 : Density plot of log2-transformed deconvolved expressions for three important genes for immune cells. Red curve represents CD4; green represents CD8A; and blue represents HLA-DQB1. Solid lines are for the immune component; dotted lines are for the stromal component; and long dashed lines are for the tumor component.

2.7 Discussion

Cellular heterogeneity is an important feature of human cancer, which strongly influences downstream analysis based on transcriptomes. Expression profiling technique has demonstrated its potential to disclose useful information for cancer prognosis. Most of the methods developed for deconvolution have their limitations, and do not include tumor infiltrating immune cells into their work, when it is of great significance to understand cellular heterogeneity in terms of the interaction between tumors and immune system.

In this work, we have presented a novel statistical method for dissecting tumor mixtures on the gene expression level, which provides a fast and accurate approach to the deconvolution of tumor specimens into two or three components. Our method allows us to estimate both cell-type-specific proportions and reconstitute patient-specific

gene expression levels at the same time with little prior information. Key features of our work that differ from previous deconvolution methods are as follows.

(i) We developed a computational tool, *DeMixT*, that deconvolves the sample contexture without any prior information on tissue proportions and tumor-specific expression; the input data are gene expression levels for tumor specimens and reference profiles. Reference profiles can be derived from historical patient data or other pipelines. We do not require reference profiles for all mixing components, and allow the information of one constituent to be unknown.

(ii) It is not necessary to incorporate any cell-type-specific reference genes or borrow any other gene subset identification methods for *DeMixT*. *DeMixT* solves the deconvolution problems with steps all through computation. We also propose a two-stage procedure to select the identifiable genes between different tissue components in different steps, and therefore improve the estimation performance.

(iii) Our model follows from the biological prospective by making use of the normality of log2-transformed expression data and linearity of raw measured data. *DeMixT* is able to deconvolve expression data from tumor specimens into both two component tissues and three component tissues. We incorporate the immune infiltration into the deconvolution model. It accurately estimates tissue-specific proportions and reconstitutes expression profiles.

(iv) In the validation of immune infiltrations in virus-associated tumors, we have shown that *DeMixT* can be applied for deconvolution of samples with immune infiltration by combining two-component and three-component *DeMixT* with clinical information.

In conclusion, we develop a statistical modeling that has unique advantages over *in silico* deconvolution of bulk tumors and helps to resolve the bottleneck arising from

sample heterogeneity in cancer genomic studies. One potential limitation of *DeMixT* is the direct inclusion of the immune infiltrating component by integrating all immune cell subtypes, which may require additional clinical information and need to be pre-processed in some case. The other concern is performing deconvolution on datasets with fewer components than the number of components it is assumed. To give solutions to this problem, we need other resources to help us to identify a reasonable number of major component tissues just as what we did in the HNSCC experiment.

In our exploration following deconvolution work, we find an important problem that may significantly affect deconvolution performance. In the genomic data, there might be a small portion of genes/probes that can contribute to deconvolution. Introducing redundant noisy genes, which are non-identifiable by deconvolution model, may prevent the unambiguous determination of proportion estimation. Moreover, because all the existing *in silico* deconvolution tools assume independently distributed expression pattern for each gene, the functional relationship among genes is also a factor to bias estimation. We are under study of gene's identifiability for deconvolution, where identifiability indicates the difference of expression patterns between different components is significantly larger than their variations within each component. Inclusion of genes with large variations of expression profile but small difference among different components is detrimental to deconvolution. Differential expression analysis based on observed tumor profiles cannot discover the true difference between hidden components. Defining a distance metric of expression profiles across multiple components is also a hard job. Therefore, it will be necessary to develop a benchmark for feature selection before deconvolution.

In our future work, we will focus more on developing statistical methods to extend the deconvolution algorithm to include more number of component tissues. It will

make it possible to give proportion estimates for each leukocyte subset of tumor infiltrating immune cells. We also want to have a deconvolution model in the future that accounts for the correlation between genes, which reflect real biological activity that affects tumor-normal interaction.

Chapter 3

Bayesian Edge Regression for Undirected Graphical Models Accounting for Biological Heterogeneity

3.1 Abstract

Current work for constructing graphical models for multivariate data does not take into account the subject specific information, which can bias the conditional independence structure in heterogeneous data. According to our discussion in the first two chapters, tumor samples are inherently heterogeneous with contaminated mixtures of normal and tumor cells. Ignoring the cellular heterogeneity in tumors and modeling the population-level genomic graphs, may inhibit the discovery of the true tumor graph, which would be attenuated towards the normal graph. In this chapter, we propose a novel edge regression model for undirected graphs, which incorporates subject-level covariates to estimate the conditional dependencies. Our model allows undirected networks to vary with the exogenous covariates and is able to borrow strength from different related graphs for estimating more robust covariate-specific graphs. Bayesian shrinkage algorithms are presented to efficiently estimate and induce sparsity for generating subject-level graphs. We demonstrate the performance of our method through simulation studies. We apply our method to several real-world datasets, including cytokine measurements from blood plasma samples from hepatocellular carcinoma (HCC) patients and normal controls.

3.2 Introduction

Graphical models, which characterize the conditional dependency structure among random variables, have gain more and more popularity in genomic studies to build the networks between different biological units, including genes and proteins. We focus on undirected graphical models, where nodes index random variables and edges represent the global conditional dependency between the variables corresponding to connected nodes (Lauritzen, 1996). A popular tool in studying undirected graph is Gaussian graphical models, which corresponds the absence of an edge to a zero entry in the *precision* (or *concentration*) matrix of multivariate Gaussian distribution, so it is also well-known as the problem of *covariance selection* (Dempster, 1972). Although a great amount of literature is provided for *covariance selection*, a standard formulation of this problem restricts observations to be drawn from the same distribution (e.g. Gaussian). However, the complexity of extraneous factors in genomic studies undermines this assumption. For example, there has been much progress in the study of group-specific graphical models to describe dependency for observations collected in distinct classes (Peterson et al., 2015; Danaher et al., 2014). Ni et al. (2017) constructed directed acyclic graphs (DAG) for generalizing those group-specific factors by investigating non-static graphs affected by extraneous factors (e.g. patient-level prognostic biomarkers), which are allowed to be categorical or continuous, in genomic studies. Ni et al. (2017) proposed the concept of *graphical regression*, which formulated the inference of graphs changed by extraneous factors with a problem of regressing the graph structure on those exogenous covariates. However, he just solved this problem in a directed acyclic graph (DAG) setting, and it is still challenging in

estimating undirected graphs because of their more complicated conditional independence properties (Ni et al., 2017).

As the primary application of graphical models in genomic analyses, co-expression or regulatory networks - genes/proteins that interact with each other - may be potentially biased by environmental factors (Luscombe et al., 2004). Our research is motivated by modern cancer research, where-in tumor heterogeneity has been shown to bias the downstream analyses after global expression profiling of tumors (Farley, 2015; Junttila and de Sauvage, 2013). The tumor samples drawn from patients exhibit heterogeneity on the cellular level, because the epithelium-derived tumor interacts with the microenvironment (normal tissue), which mostly comprises non-cancer cells and surrounding stroma, during the development and progression of cancer. Those non-cancer cells contaminate gene expression profiles and add noises to detected molecular characteristics for solid tumors. We can model the transcriptional profile of a single clinically derived tumor sample as a mixture of its constituting cancerous 'tumor' and non-cancerous 'normal' component (Heppner and Miller, 1983; Liotta and Petricoin, 2000). Tumor purity, which is the proportion of cancerous tissues, measures the degree of normal contamination and varies widely among solid tumor samples pre-selected for genomic analyses. It adds major confounding factors resulting from tumor-normal interactions to bias the final results (West et al., 2010; Bachtiary et al., 2006). For assays run on solid tumor samples, the sample will often be a mixture of normal and tumor cells that cannot practically be separated from each other. By estimating and accounting for the tumor purity for each sample we can gain power for detecting tumor-normal differences. In other settings, heterogeneity can also be a factor. Although a majority of deconvolution methods are provided for estimating tumor purity from mixed tumor samples, they cannot accurately recover co-expression

patterns for different compartments in solid tumors and maintain the dependency structure across molecular units due to their assumption of independently expressed genes (Yadav and De, 2015). We are interested in studying the common and differential dependency structure within tumor and normal component from clinically derived tumor samples, where 100% pure tumors are rarely collected. Current approaches for this study just consider those mixed tumors with a population of reference normal as two different classes of samples, which cannot recover the network for pure tumors. In this sense, *graphical regression* provides us with coherent inferential framework on how the structure of tumor network varies with tumor purity, and finally helps us to recover a pure 'normal graph' (network structure for non-cancerous tissues) and a pure 'tumor graph' (network structure for cancerous tissues) from those learnt subject-level networks by substituting tumor purity with 0% and 100%.

3.3 Existing methods for undirected graphical models

The use of *G-Wishart* distribution as a conjugate prior for the *precision* matrix and penalized likelihood (Friedman et al., 2008; Huang et al., 2006) for the sparsity control are the two major traditional means of inference for Gaussian graphical models. Recently developed regularization methods through *neighborhood selection*, which implement penalized regression per node over all the other nodes, provide us with different approaches for inference of undirected graphs (Meinshausen and Bühlmann, 2006). Several methods using a joint sparse regression and Bayesian shrinkage prior have been suggested for new benefits based on *neighborhood selection* (Leday et al., 2015; Peng et al., 2012). There also have been several attempts to model undirected graphs with a similar flavor as *graphical regression*. Hoff and Niu (2012) proposed

a covariance regression model by regressing covariance matrix over explanatory variables for factor analysis. Zhou et al. (2010) and Kolar and Xing (2009) developed dynamic undirected graph models varying with time. Cheng et al. (2014) modeled multivariate binary data using an Ising model to study the change of dependency with covariates. Liu et al. (2010) proposed an algorithm *Graph-optimized classification and regression trees* to partition the covariate space and estimate the graph within each partition subspace. However, Cheng et al. (2014) reported that this model is lack of interpretation of the relationship between graphical models and covariates, and it is unstable since graphs constructed for close covariates are not necessarily similar. A machine learning method proposed by Kolar et al. (2010) applied a penalized kernel smoothing approach and allowed the *precision* matrix to change with covariates, but this method is limited by ignoring the intrinsic symmetry of elements in *precision* matrix, which may result in contradiction for neighborhood selection and subsequent interpretation.

In this work, we developed a Bayesian edge regression (ER) approach for undirected graphical models. Similar in spirit to Ni et al. (2017), we define an edge-specific *conditional precision function* to allow the edge strength for undirected graphical models to vary with the exogenous covariates. This function is linked with the estimation of element in the *precision* matrix through a joint regression model, hence constraining the elements corresponding to the same pair of nodes to be exactly same and guaranteeing the symmetry of estimated *precision* matrix. This subsequently helps us to better interpret the relationship between the graph structure and covariates and also allows predicting graph structure for new observations. The use of adaptive Bayesian shrinkage prior induces the local shrinkage of edge strength with a global shrinkage of the regularizing parameter over exogenous covariates across different edges. In doing

so, we impose two-level shrinkage on the edge strength and exogenous covariates. We combine a scheme of edge selection that allows coherent multiplicity controls of the expected global Bayesian false discovery rate (FDR). We discuss the parameterization of precision function and provide several parameterization solutions depending on the scientific and inferential context. Finally we illustrate the application of our model to infer networks in heterogeneous tumor samples and group specific observations through simulations and case studies using cancer genomic and proteomic data. We apply this method to a prostate cancer data set and a liver cancer cytokine study to estimate blood plasma cytokine networks induced by hepatocellular carcinoma and those from normal controls while accounting for biological heterogeneity.

The rest of chapter is structured as follows. In Section 3.4, we provide a formal description of edge regression with several theoretical properties for undirected graphical models. Then we present our models with sampling scheme and posterior inference technique. We present our simulation studies in Section 3.5 and include case studies on prostate cancer samples and blood plasma samples from hepatocellular carcinoma with application of our method in Section 3.6. Section 3.7 concludes this chapter with discussion.

3.4 Methods

3.4.1 Edge regression

A graphical model for a random p-vector $\boldsymbol{Y}$ is defined by a tuple $\mathcal{G}_{\mathcal{Y}} = \{G, \mathcal{P}(\boldsymbol{Y})\}$, where G is a graph and $\mathcal{P}(\boldsymbol{Y})$ denotes its associated distribution. $G = (V, E)$ represents conditional independence structure among random variables by specifying a set

of nodes $V = 1, 2, 3, \cdots, p$ and a set of edges $E \in V \times V$. Each node in the graph G corresponds to a random variable in $\boldsymbol{Y}$. In an undirected graph, we have undirected edges E, where $(i, j) \in E$ if and only if $(j, i) \in E$. For example, a Gaussian graphical model is defined by assuming $\mathcal{P}(Y)$ is a Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$.

$$\boldsymbol{Y_n} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}), n = 1, \cdots, N \tag{3.1}$$

, where $\boldsymbol{Y_n}$ is the observed data and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{p \times p}$ is the inverse covariance matrix (a.k.a precision matrix or concentration matrix). In Gaussian graphical model, $\boldsymbol{\Omega}$ is a symmetric positive definite matrix and we denote the precision matrix by $(\omega^{ij})_{p \times p}$. If $\omega^{ij} = 0$, then the random variable i and j are conditionally independent given all the other variables of $\boldsymbol{Y}$, which indicates there is no edge in G between node i and node j. Therefore the parameterization of precision matrix $\boldsymbol{\Omega}$ can be bridged with the construction of the conditional independence structure in the graph G. This is well-known as the *covariance selection* model.

In our proposed edge regression model, given another q-dimension random vector $\boldsymbol{X} = (x_1, \cdots, x_q)^T$, we consider $\mathcal{G}_{\mathcal{Y}}(\boldsymbol{X}) = \{G(\boldsymbol{X}), \mathcal{P}(\boldsymbol{Y}|\boldsymbol{X})\}$ and the precision matrix for each observation $\boldsymbol{Y_n}$ given $\boldsymbol{X} = \boldsymbol{x_n}$ is a function of $\boldsymbol{X}$, reflecting that the conditional independence structure is able to vary from observation to observation over different realizations of $\boldsymbol{X}$. In the following discussion, we use the term exogenous covariates to define $\boldsymbol{X}$. We denote the precision matrix dependent on $\boldsymbol{X}$ through $\Omega(\boldsymbol{x})$ so its element $\omega^{ij}(\boldsymbol{X})$. $\omega^{ij}(\boldsymbol{X})$ is specific to the edge (i, j) (equivalently, (j, i)) and can be allowed to be a linear relationship. We formally state a lemma of edge regression for *covariance selection* problem.

Lemma 1 (FUNCTIONAL COVARIANCE SELECTION RULE) Assuming $\boldsymbol{Y}$ *has a multivariate Gaussian distribution given exogenous covariates* $\boldsymbol{X}$ *with a precision matrix* $\Omega(\boldsymbol{X})$. $Y^i \perp\!\!\!\perp Y^j | \boldsymbol{Y}^{-(\boldsymbol{i},\boldsymbol{j})}, \boldsymbol{X} \Leftrightarrow \omega^{ij}(\boldsymbol{X}) = 0$

Lemma 1 follows from the covariance selection rule when a set of exogenous covariates $\boldsymbol{X}$ is given. Following Ni et al. (2017), these special cases of edge regression are provided:

(1) If $\boldsymbol{X} = \emptyset$, then edge regression model reduces to the case of the ordinary undirected graphical model.

(2) If $\boldsymbol{X}$ is a set of discrete covariates (e.g., binary/categorical), then edge regression model reduces to the problem of estimating group-specific graphical models.

3.4.2 Regression model for undirected graphs

In this section, we introduce a sparse regression model for implementation of edge regression. From now on we assume the $\boldsymbol{\mu} = 0$ for simplicity. Denote the partial correlation between random variable Y^i and Y^j by $\rho^{ij} (1 \leq i \neq j \leq p)$, where $\rho^{ij} = -\frac{\omega^{ij}}{\sqrt{\omega^{ii}\omega^{jj}}}$. Hence, from the covariance selection rule, the edge $(i, j) \in E$ is equivalent to the partial correlation $\rho_{ij} \neq 0$. A well-known lemma implies that when y^i $(1 \leq i \leq p)$ is expressed in a linear regression form of $\sum_{j \neq i} \gamma^{ij} y^j + \epsilon_i$, $\gamma^{ij} = -(\omega^{ij}/\omega^{ii})$ and ρ^{ij} can be represented as $sign(\gamma^{ij})\sqrt{(\gamma^{ij}\gamma^{ji})}$. We can extend this lemma to a case of edge regression by including the exogenous covariates $\boldsymbol{X}$ into the regression method.

Lemma 2 For $1 \leq i \leq p$*, considering predicting* y^i *from other variables* y^{-i} *given exogenous covariates* $\boldsymbol{X} = \boldsymbol{x}$ *with a varying-coefficient model, we have* $y^i = \sum_{j \neq i} \gamma^{ij}(\boldsymbol{x}) y^j + \epsilon_i$*, such that* ϵ_i *is uncorrelated with* y^{-i} *given* $\boldsymbol{X} = \boldsymbol{x}$ *if and only if the optimal prediction rule gives* $\gamma^{ij}(\boldsymbol{x}) = -\frac{\omega^{ij}(\boldsymbol{x})}{\omega^{ii}(\boldsymbol{x})} = \rho^{ij}(\boldsymbol{x}) \sqrt{\frac{\omega^{jj}(\boldsymbol{x})}{\omega^{ii}(\boldsymbol{x})}}$*, where* $\omega^{ii}(\boldsymbol{x})$ *and* $\omega^{ij}(\boldsymbol{x})$ *respectively correspond to the off-diagonal and on-diagonal element of* $\Omega(\boldsymbol{x})$*. Hence,* $\rho^{ij}(\boldsymbol{x}) = sign(\gamma^{ij}(\boldsymbol{x})) \times \sqrt{\gamma^{ij}(\boldsymbol{x}) \gamma^{ji}(\boldsymbol{x})}$*. Additionally,* $var(\epsilon_i) = 1/\omega^{ii}(\boldsymbol{x})$*.* $\gamma^{ij}(.)$ *is a conditional precision function (CPF) that defines* ρ^{ij} *through* $\boldsymbol{X}$*.*

Lemma 2 is also self-evident when the partial correlation is calculated given $\boldsymbol{X}$. From Lemma 2, $\boldsymbol{X}$ changes the partial correlation ρ^{ij} as well as the regression coefficients of y^i over y^j through the function $\gamma^{ij}(.)$. In this sense, $\gamma^{ij}(.)$ can be fitted to characterize the conditional dependency structure for a subject-level graph given $\boldsymbol{X}$. Under this setting, the covariance selection problem for a subject-level graph can be converted to a feature selection problem for regression with varying coefficients (Hastie and Tibshirani, 1993), i.e., the sparsity structure of undirected graph can be learnt through a sparse regression. When we apply edge regression, the shrinkage of $\gamma^{ij}(\boldsymbol{x})$ to exact zero can be realized through the shrinkage of each coefficient in the function of $\gamma^{ij}(.)$. Although the whole precision matrix is currently allowed to vary with $\boldsymbol{X}$, for model parsimony and simplicity, we assume the on-diagonal element ω^{ii} constant across different $\boldsymbol{X}$ due to our primary interest on the relationship between the edge structure (determined by off-diagonal element ω^{ij}) and $\boldsymbol{X}$. Hence, the study of how exogenous covariates affect edge selection is equivalent to learning how the sparsity structure of off-diagonal element in precision matrix varies with $\boldsymbol{X}$.

3.4.3 Parameterization of the conditional precision function

In Lemma 2, we define the *conditional precision function*, which can be parameterized with linear or nonlinear form. Suppose we have a set of exogenous covariates $\boldsymbol{X}$, which can include discrete and continuous covariates. According to our assumption, $\omega^{ii}(\boldsymbol{x}) = \omega^{ii}$. With $\gamma^{ij}(\boldsymbol{x}) = -\frac{\omega^{ij}(\boldsymbol{x})}{\omega^{ii}}$, $\gamma^{ij}(.)$ can be functionally determined by $\omega^{ij}(.)$, which is restricted to be equal to $\omega^{ji}(.)$ in precision matrix. If we use a linear function to model the relationship between the partial correlations and exogenous covariates, $\omega^{ij}(.)$ can be parameterized through $\boldsymbol{X}$:

$$\omega^{ij}(\boldsymbol{X}) = \sum_{s=1}^{q} \beta_s^{ij} X_s \tag{3.2}$$

, where β_s^{ij} is the effect of discrete or categorical variable X_s on the edge (i, j). Although we parameterize the *conditional precision function* specific to our problem in the following analysis, we discuss several instances for parameterization.

(1) The functional relationship between ω^{ij} and $\boldsymbol{X}$ is equal to that between ρ^{ij} and $\boldsymbol{X}$.

(2) The parameterization should be adapted to real background of $\boldsymbol{X}$, because the regularization method for covariate selection can be affected by the manner of parameterization. For example, in our case study of tumor heterogeneity problem, a cell mean model is preferred for respectively shrinking edges in both pure tumor and normal graphs.

(3) For categorical covariates $\boldsymbol{Z}$, edge regression is a group-specific model. Dummy coding of $\boldsymbol{Z}$ with interaction terms between different groups can borrow strength for estimation.

For each single regression for edge (i, j), the *conditional precision function* can be considered as a variant of *varying coefficient model* (Hastie and Tibshirani, 1993). In this work, we only discuss the estimation problem in the setting of a linear expression for the *conditional precision function*. In the section of case studies, we present our parameterization strategy for several specific problems.

3.4.4 Bayesian adaptive shrinkage

As previously discussed, the sparseness of estimated precision matrix given exogenous covariates corresponds to the absence of edges in a subject-level graph. It has been shown that most genomic graphs can be truly sparse. In other words, the estimated precision matrix $\Omega(\boldsymbol{X})$ is expected to be sparse enough, so ω^{ij} is shrunk if there is no evidence for it to be non-zero. We adopt a Bayesian approach to combine local regularization for each regression parameter with global shrinkage of the regularizing parameters across edges for each exogenous covariate (O'Hara et al., 2009). Edges with small magnitude of elements in precision matrix after shrinkage will be threshed out. Assuming a linear model for *conditional precision function*, $\omega^{ij}(\boldsymbol{X}) = \sum_{s=1}^{q} \beta_s^{ij} X_s$, we impose local shrinkage model with normal-gamma prior to shrink β_s^{ij} with weak evidence to be non-zero towards 0. The shrinkage of β_s^{ij} can lead to the shrinkage of $\beta_s^{ij}\boldsymbol{x_s}$. ρ^{ij} given $\boldsymbol{X}$ with most of $\beta_s^{ij}\boldsymbol{x_s}$ shrunk towards zero will also be shrunk towards zero at the same time. Then we set a standardized threshold to thresh out small ρ^{ij} given $\boldsymbol{X}$ to accomplish the edge selection. Normal-gamma prior benefits us by only extreme shrinkage of "small" coefficients but weak shrinkage of "large" coefficients (Griffin et al., 2010). The existence of closed-form posterior probability of regularization parameters for normal-gamma prior also makes it convenient to implement a Gibbs sampling method.

By regressing Y^i over $\boldsymbol{Y^{-i}}$ given $\boldsymbol{X}$, we can write our model as:

$$\begin{aligned} Y^i &= \sum_{j \neq i} \gamma^{ij}(\boldsymbol{X}) Y^j + \epsilon^i, i = 1, \cdots, p \\ \gamma^{ij}(\boldsymbol{X}) &= -\frac{\omega^{ij}(\boldsymbol{X})}{\omega^{ii}}; \epsilon^i \sim N(0, \frac{1}{\omega^{ii}}) \end{aligned} \tag{3.3}$$

Since $\omega^{ij}(.)$ is the off-diagonal element of precision matrix corresponding to vertex i and vertex j, we have $\omega^{ji}(.) = \omega^{ij}(.)$. Hence we can coerce these two functions to have the same formula in the sampling scheme. Consequently, we have $\beta_s^{ij} = \beta_s^{ji}$ for every $i \neq j$.

We rewrite the full conditional probability of Y^i as:

$$Y^i | \boldsymbol{Y^{-i}}, \{\boldsymbol{\beta_s^{i,-i}}\}_{s=1}^q, \omega^{i,i}, \{X_s\}_{s=1}^q \sim N(-\frac{\sum_{j \neq i}^p \sum_{s=1}^q \beta_s^{ij} X_s y^j}{\omega^{ii}}, \frac{1}{\omega^{ii}}) \tag{3.4}$$

Normal-Gamma prior The normal-gamma prior for edge regression is given in a hierarchical form:

$$\begin{aligned} \beta_s^{ij} &\sim N(0, \psi_s^{ij}); \pi(\omega^{ii}) \propto 1 \\ \psi_s^{ij} &\sim Gamma(\lambda_s, 1/(2\gamma^2)) \end{aligned} \tag{3.5}$$

By assuming ω^{ii} not varying with exogenous covariates, the CPF $\gamma^{ij}(x) = -\frac{\omega^{ij}(x)}{\omega^{ii}} = -\frac{\sum_{s=1}^q \beta_s^{ij} X_s}{\omega^{ii}} = \sum_{s=1}^q -\frac{\beta_s^{ij}}{\omega^{ii}} X_s$ still satisfies a linear function. For each β_s^{ij} in edge regression, a regularization parameter ψ_s^{ij} of normal prior is set to locally shrink each coefficient. For different ψ_s^{ij} regularizing the same covariate X_s across different edges, we have the same hyper-parameter λ_s, γ of gamma prior to globally control them. The scale parameter γ is set to be same across all the covariates. For ω^{ii}, which controls the variance parameter in the neighborhood selection model, we choose a vague prior such that $\pi(\omega^{ii}) \propto 1$ as done by Griffin et al. (2010) for our following discussion. If

ω^{ii} is given with a conjugate prior $Gamma(a^*, b^*)$, the full conditional distribution for ω^{ii} keeps the same form, so our sampling scheme can still be implemented by a Gibbs step. A graphical representation of the hierarchical formulation corresponding to normal-gamma prior is shown in Figure 3.1.

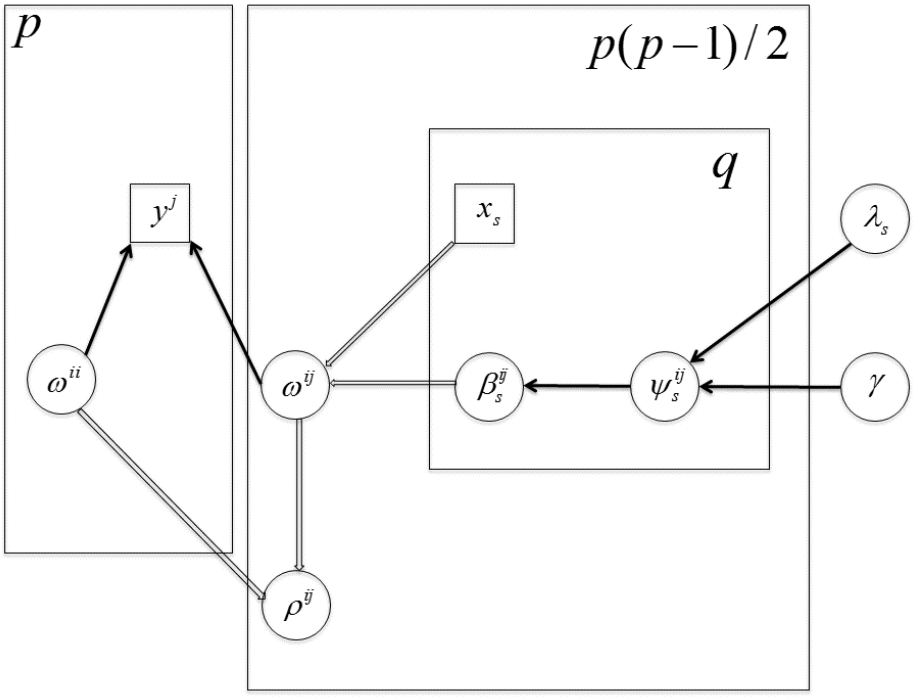


Figure 3.1 : A graphical representation of edge regression with normal-gamma prior. Single arrows are probabilistic edges; double arrows are deterministic edges; squares are observed data; circles are random variables. The total number of instances of each variable that is enclosed in the same plate is given by the constant in the corner of that plate. ρ^{ij} is the partial correlation for edge (i, j) that we want to finally get.

Sampling scheme We follow the scheme in Griffin et al. (2010) to sample λ_s and γ simultaneously by specifying exponential and inverse-gamma hyper priors. According to this hierarchy of normal-gamma prior, we implement a block Metropolis-within-Gibbs sampling scheme to update each parameter sequentially. A summary

of notation we use and details of our derivation can be found in Appendix B.1 and B.2.

- Update β^{ij} for every pair $(i,j), i<j$

  For $\boldsymbol{\beta^{ij}} = \{\beta_s^{ij}\}^S$ of any given pair of vertex (i,j), the full conditional distribution follows a multivariate Gaussian distribution with mean

$$\widetilde{\boldsymbol{\mu}}^{\boldsymbol{ij}} = -(\boldsymbol{X}^T \boldsymbol{S_1} \boldsymbol{X} + (\boldsymbol{\psi}^{\boldsymbol{ij}})^{-1})^{-1} \boldsymbol{X}^T \boldsymbol{S_2} \tag{3.6}$$

  and variance

$$\widetilde{\boldsymbol{\Sigma}}^{\boldsymbol{ij}} = (\boldsymbol{X}^T \boldsymbol{S_1} \boldsymbol{X} + (\boldsymbol{\psi}^{\boldsymbol{ij}})^{-1})^{-1} \tag{3.7}$$

  where $\boldsymbol{X} = (\boldsymbol{X_1^T}, \boldsymbol{X_2^T}, \cdots, \boldsymbol{X_N^T})^{\boldsymbol{T}}$ is an $N \times q$ matrix with each row describing observed$X$ for each sample. $\boldsymbol{S_1} = diag(S_{1,1}, S_{1,2}, \cdots, S_{1,N})$ is an $N \times N$-vector and $\boldsymbol{S_2} = \{S_{2,1}, S_{2,2}, \cdots, S_{2,N}\}^T$ is an N-dimensional vector. $S_{1,n}$ and $S_{2,n}$ are given as:

$$\begin{aligned} S_{1,n} &= \frac{(Y_n^j)^2}{\omega^{ii}} + \frac{(Y_n^i)^2}{\omega^{jj}} \\ S_{2,n} &= 2Y_n^i Y_n^j + \boldsymbol{X_n}^T \boldsymbol{\beta^{i,-j}} \boldsymbol{Y_n^{-(i,j)}} \frac{\boldsymbol{Y_n^j}}{\omega^{ii}} + \boldsymbol{X_n}^T \boldsymbol{\beta^{j,-i}} \boldsymbol{Y_n^{-(j,i)}} \frac{\boldsymbol{Y_n^i}}{\omega^{jj}} \end{aligned} \tag{3.8}$$

- Update $\omega^{ii}, i = 1, 2, \cdots, p$

  The full conditional distribution of ω^{ii} is:

$$GIG(\frac{n}{2}+1, \sum_{n=1}^{N}(Y_n^i)^2, diag(\boldsymbol{X\beta^{i\cdot}}(\boldsymbol{Y^{-i}})^T) diag(\boldsymbol{X\beta^{i\cdot}}(\boldsymbol{Y^{-i}})^T)^T) \tag{3.9}$$

  where $GIG(m,a,b)$ is the Generalized Inverse Gaussian distribution. It has the density

$$f(x) = \frac{(a/b)^{m/2}}{2K_m(\sqrt{ab})} x^{(m-1)} e^{-(ax+b/x)/2} \tag{3.10}$$

- Update ψ_s^{ij}

  ψ_s^{ij} for the edge (i, j) and s-covariate can be effectively updated in a block since their full conditional distributions are independent. The full conditional distribution also follows a Generalized Inverse Gaussian distribution with:

$$GIG(\lambda_s - \frac{1}{2}, 1/\gamma_s^2, (\beta_s^{ij})^2) \tag{3.11}$$

- Update hyper-parameters of the normal-gamma prior

  We assigned prior $\pi(\lambda_s) = exp(1)$ for the shape parameter λ_s, then the full conditional λ_s is proportional to:

$$\propto \pi(\lambda_s) \frac{1}{(2\gamma^2)^{\frac{p(p-1)}{2}\lambda_s} (\Gamma(\lambda_s))^{\frac{p(p-1)}{2}}} (\prod_{i \neq j} \psi_s^{ij})^{\lambda_s} \tag{3.12}$$

  For the scale parameter γ, we specify a prior $\sum_s \lambda_s \gamma^2 \sim Ga(2, \sum M_s))$. M_s is a hyper-parameter to approximately control the scale of $\lambda_s \gamma^2$ for the s-th covariate. The calculation of M_s is discussed in our Appendix B.2 for specific problems. We have:

$$\gamma^{-2} \sim Ga(2 + qp(p-1)\lambda_s/2, \sum_s M_s/(2\sum_s \lambda_s) + \frac{1}{2} \sum_s \sum_{i \neq j} \psi_s^{ij}) \tag{3.13}$$

3.4.5 Posterior inference and thresholding

An algorithm demonstration for our MCMC sampling scheme for posterior inference is given in Algorithm 2. To carry out the edge selection after the imple-

mentation of MCMC, we first need to calculate a posterior probability of inclusion (PPI) for each edge. Given the observed data of exogenous covariates X, for each pair of nodes (i, j), $\rho_n^{ij,l}$ given $X = x_{(n)}$ can be calculated in each iteration of MCMC. The approach of adaptive shrinkage prior shrinks covariates towards zero but not exact zero. So we still need to set a threshold κ to indicate the inclusion of each edge in each iteration, such that $(i, j) \in E$ if $|\rho_n^{i,j}| > \kappa$ (Hoti and Sillanpää, 2006). The marginal posterior probability of inclusion (PPI) is calculated by $Pr((i, j) \in E|y, x_{(n)}) = \sum_l^{L_T} I(|\rho_n^{ij,l}| > \kappa)/L_T$, where L_T is the thinned posterior sample size after bur-in period. It is the posterior probability that the absolute effective size exceeds the given threshold. The choice of κ is subjective and should be determined based on the specific context. Too large or too small κ can lead to high false positive rate or false negative rate. In this work, we choose $\kappa = 0.1$, because it results in reasonable discovery of true positives and negatives.

Given our defined marginal posterior probability of inclusion, we select a set of edges in a way by controlling the expected Bayesian FDR at level α. This rule was first proposed for detecting differentially expressed genes but we found it useful in the context of edge structure discovery here by considering both statistical and practical significance (Morris et al., 2008). If we let $q_n^{i,j}$ denote the quantity $1 - Pr((i, j) \in E|y, x_{(n)})$, it can be considered an estimate of local FDR for selecting edge (i, j). Following the rule of controlling global Bayesian FDR, we select the set of edges $E_n = \{(i, j) : q_n^{i,j} < \phi_{n,\alpha}\}$ for each sample n. Our FDR controlling procedure works as follow.

1. Sort $\{q_n^{i,j}, (i, j) \in E\}$ in ascending order to obtain $\{q_n^{(t)}, t = 1, \cdots, |E|\}$

2. For a given α, find the largest $t*$ such that $(t*)^{-1}\sum_{t=1}^{t*} q_n^{(t)} < \alpha$

3. Set $\phi_{n,\alpha} = q_n^{(t*)}$, and select edges with $q_n^{i,j} < \phi_{n,\alpha}$

We control the expected Bayesian FDR at level α, which implies that on average $\leq 100\alpha\%$ of the edges in the set E_n will result in false positives. This choice of cutoff α can depend on cases because stringent α leads to fewer false positive but also fewer true positive. In our current work of simulation, we found that $\alpha = 0.1$ resulted a reasonable true positive rate, false positive rate. We will also compute the average area under the curve (AUC) of the receiver operating characteristic (ROC) curve to examine our model performance under different thresholds of κ and α.

Algorithm 1 MCMC sampling scheme under normal-gamma prior for edge regression

Initialize:
  $\{\beta_s^{ij}\}_{s\in S}^{1<i\neq j<p}$, $\{\omega^{ii}\}_{i=1}^{p}$, $\{\psi_s^{ij}\}_{s\in S}^{1<i\neq j<p}$, $\{\lambda_s\}_{s\in S}$, $\{\gamma_s\}_{s\in S}$
for iteration $l = B+1, \cdots, L$, (B is the burn-in period) **do**
  a. update $\beta_s^{ij,l}$, $\omega^{ii,l}$, $\psi_s^{ij,l}$ by a Gibbs step
  b. update λ_s and γ_s by a Metropolis-Hasings step
  for each sample $n = 1, \cdots, N$ **do**
    calculate $\omega^{ij,l}(x_{(n)})$ from $\beta_s^{ij,l}$ and $x_{(n)}$
    calculate $\rho^{ij,l}(x_{(n)})$ from $\omega^{ij,l}(x)$ and $\omega^{ii,l}$ given $X = x_{(n)}$
  end for
end for
Output thinned posterior samples of $\rho^{ij}(x_{(n)})$

3.5 Simulations

In this section, we include one simulation experiment to highlight the performance of our model under the genomic background for the tumor heterogeneity problem. We compare our proposed method with two approaches to estimating multiple graphical

models, fused graphical lasso and group graphical lasso, in terms of edge selection (Danaher et al., 2014). In our simulation, we have all continuous exogenous covariates. This case corresponds to an important application of edge regression model to tumor heterogeneity problem we emphasize in the section of introduction. We elaborate more on the problem of tumor heterogeneity in the section of simulation and application. We also include another simulation, which is a special case of edge regression when each subject-level graph is group specific. For each simulation, we run $20,000$ MCMC iterations, in which first $10,000$ iterations are discarded as a "burn-in" period, and thin out the chain using every 10-th sample.

3.5.1 Case I: continuous exogenous covariates

In this simulation, we consider exogenous covariates $\boldsymbol{X}$ to be continuous and simulate data in a way from tumor sample deconvolution problem. We include 20 nodes to represent 20 genes. We generate observed tumor expressions by log2-transforming expressions mixed on the raw level from simulated tumor and normal components, where log2-transformed data still satisfy a normal distribution. According to Chapter 2, observed expressions from clinically derived tumor samples are assumed to be a linear mixture of the expressions from pure normal and pure tumor components before log2-transformation of gene expression data. It follows that,

$$2^{Y_n} = (1 - \pi_n)2^{N_n} + \pi_n 2^{T_n} \tag{3.14}$$

where $\boldsymbol{N_n} \sim \mathcal{N}(\boldsymbol{\mu_N}, \boldsymbol{\Omega_N^{-1}})$ and $\boldsymbol{T_n} \sim \mathcal{N}(\boldsymbol{\mu_T}, \boldsymbol{\Omega_T^{-1}})$. $\pi_n \in [0, 1]$ is the measured tumor purity for sample n.

For simplicity, we set $\boldsymbol{\mu_N} = \mathbf{0}$ and $\boldsymbol{\mu_T} = \mathbf{0}$ in our simulation. We also generate

$\boldsymbol{N}'_{\boldsymbol{n}} \sim \mathcal{N}(\mathbf{o}, \boldsymbol{\Omega}_{\boldsymbol{N}}^{-\mathbf{1}})$ as a reference group for normal component. In the current study of cancer pathway, expressions from normal samples, i.e. $\boldsymbol{N}'$, and clinically derived tumor samples, i.e. $\boldsymbol{Y}$, are analyzed separately. A most recent fashion construct graphs from these tumor and normal samples by considering they are group specific. All these attempts discover graphs from mixed expressions of $\boldsymbol{Y}$ but fail to learn the structure for $\boldsymbol{T}$, which represents the cancerous tumor component after removal of normal contamination. In our following simulation, we provide two simulations with different set-ups of precision matrix.

Simulation 1. $\boldsymbol{\Omega_T}$, where off-diagonal elements $\omega_1^{i,i+2} = \omega_1^{i+2,i}$ uniformly sampled from $[-0.5, -0.3] \cup [0.3, 0.5]$ for $i = 1, \cdots, 18$. $\boldsymbol{\Omega_N}$, where off-diagonal elements $\omega_1^{i,i+1} = \omega_1^{i+1,i}$ uniformly sampled from $[-0.5, -0.3] \cup [0.3, 0.5]$ for $i = 1, \cdots, 19$. For both $\boldsymbol{\Omega_T}$ and $\boldsymbol{\Omega_N}$, all the diagonal elements are 1 and all the other elements are left with zero.

Simulation 2. $\boldsymbol{\Omega_T}$, where diagonal elements $\omega_1^{i,i} = 1$ for $i = 1, \cdots, 20$ and off-diagonal elements $\omega_1^{i,i+1} = \omega_1^{i+1,i} = 0.5$ for $i = 1, \cdots, 19$, $\omega_1^{i,i+2} = \omega_1^{i+2,i} = 0.4$ for $i = 1, \cdots, 18$. All the other elements are left with zero. $\boldsymbol{\Omega_N}$, where we remove 30 edges randomly from $\boldsymbol{\Omega_T}$ by substituting these 30 nonzero elements with zero and randomly add 30 edges to $\boldsymbol{\Omega_T}$ by substituting these 30 zero elements with values uniformly sampled from $[-0.6, -0.4] \cup [0.4, 0.6]$. $\boldsymbol{\Omega_T}$ is a $AR(2)$ model. To ensure $\boldsymbol{\Omega_N}$ to be positive definite, we divide each off-diagonal element by 1.5 times the sum of the absolute value of all the off-diagonal elements in its row. Then we average the transformed matrix with its transpose to guarantee it is symmetric.

In our set-up of *Simulation 1*, $\boldsymbol{\Omega_T}$ and $\boldsymbol{\Omega_N}$ are truly sparse with just 18 and 19 edges. They do not have any overlapping edge and weak signal. We simulate reference normal samples of size $N_{N'} = 50$ and mixed tumor samples of size $N_Y = 150$

with $\{\pi\}_{n=1}^{150}$ generated from an arithmetic sequence from 0.01 to 0.99. In *Simulation 2*, we allow $\boldsymbol{\Omega_T}$ and $\boldsymbol{\Omega_N}$ to have 7 overlapping edges, and $\boldsymbol{\Omega_N}$ has weak signals. We simulate reference normal samples of size $N_{N'} = 100$ and mixed tumor samples of size $N_Y = 200$ with the same way of generating π_n.

We use fused and group graphical lassos that are available in *R* package *JGL* as the comparison method by considering $\boldsymbol{Y}$ and $\boldsymbol{N}'$ group specific. So mixed samples are analyzed by these competing methods to learn the structure for pure component $\boldsymbol{T}$ and the structure for pure component $\boldsymbol{N}$ with reference normal. Then we follow the same procedure recommended in Danaher et al. (2014) by searching over a grid of possible values for tuning parameter λ_1 and λ_2. We then choose the combination that minimizes the approximate $AIC(\lambda_1, \lambda_2)$ score. In the context of this problem, we adopt the following parameterization for our *conditional precision function*:

$$\omega^{ij}(\pi) = \beta^{ij}(1-\pi) + \alpha^{ij}(\pi) \tag{3.15}$$

We parameterize the relationship between precision element and π, the purity of component $\boldsymbol{T}$, in a similar way with cell mean models. With purity of component $\boldsymbol{N}$ being $1-\pi$, the population-level graph of pure component $\boldsymbol{T}$ and $\boldsymbol{N}$ can be given by β^{ij} and α^{ij} straightforward because:

$$\omega^{ij}(\pi) = \begin{cases} \beta^{ij}, & \pi = 0 \\ \alpha^{ij}, & \pi = 1 \end{cases} \tag{3.16}$$

Hence the shrinkage over β^{ij} and α^{ij} makes shrinkage of edge strength in graph for $\boldsymbol{T}$ and $\boldsymbol{N}$ more smooth. The hyper-parameter M_N and M_T is set by calculating $\hat{\Omega}_{MLE}$ respectively for samples with tumor purity below 0.5 and above 0.5. This derivation is intuitive, since only the scale of M is of primary importance for a as a

hyper-parameter. We implement these methods across 100 simulated data sets for the first simulation, 77 simulated data sets for the second simulation. We evaluate the accuracy of estimating the graph structure in terms of the true positive rate (TPR), false positive rate (FPR) and the area under the ROC curve (AUC). For both two graphical lasso methods and our method, we all have two threshold parameters. (κ and α; λ_1 and λ_2). Hence, we calculate a bivariate AUC (McGuffey and et al, 2017) by varying both two parameters at the same time. TPR and FPR are provided for $\kappa = 0.1$ and $\alpha = 0.1$. Results for those two simulations above are given in Table 3.1 and 3.2. In *Simulation 1*, since there is no weak signal in both normal and tumor graphs, all those three methods can achieve high TPR and bAUC in learning structure of normal graph. The edge regression method is able to discover fewer false edges to give FPR lower than 0.1 no matter for normal or tumor graph. It outperforms the other lasso methods in terms of all these three measures for the edge selection of tumor graph, for which we do not provide any reference information. In *Simulation 2*, the small magnitude of elements in precision matrix makes the structure of normal graph more difficult to be learnt. Our method still keeps a low FPR, even though the TPR is decreased. The edge regression method is more powerful in learning structure of tumor graph and our bAUC is higher than the other two methods across all these simulation settings. The corresponding bivariate ROC curves are provided in Figure 3.2. A bivariate ROC curve describes how the average true positive changes with average false positive through controlling two thresholds. In those lasso methods, the *fused lasso* and the *group lasso* penalty are added to respectively penalize differences between groups and elements across all precision matrices in addition to *lasso* penalty applied to the element of precision matrices (Danaher et al., 2014). With two discrimination thresholds, even though for one

false positive value, there could be many different corresponding true positive values. We calculate the averaged (or maximum) false positive corresponding to each true positive value. Hence, the ROC curve might not be monotone.

Table 3.1 : Results of edge selection for *Simulation 1* in terms of TPR, FPR and bAUC. The numbers are averaged across 100 simulated sets and the standard deviations are given within the parentheses.

Method		Normal	Tumor	Overall
Fused graphical lasso				
	TPR	0.993 (0.018)	0.812 (0.117)	0.902 (0.060)
	FPR	0.510 (0.075)	0.355 (0.093)	0.433 (0.080)
	bAUC	0.948 (0.010)	0.674 (0.060)	0.811 (0.029)
Group graphical lasso				
	TPR	0.993 (0.018)	0.824 (0.103)	0.908 (0.053)
	FPR	0.522 (0.086)	0.366 (0.100)	0.444 (0.090)
	bAUC	0.939 (0.014)	0.769 (0.051)	0.854 (0.026)
Bayesian edge regression				
	TPR	0.982 (0.028)	0.838 (0.095)	0.910 (0.049)
	FPR	0.094 (0.027)	0.083 (0.027)	0.089 (0.019)
	bAUC	0.947 (0.011)	0.916 (0.030)	0.932 (0.016)

Table 3.2 : Results of edge selection for *Simulation 2* in terms of TPR, FPR and bAUC. The numbers are averaged across 77 simulated sets and the standard deviations are given within the parentheses.

Method		Normal	Tumor	Overall
Fused graphical lasso				
	TPR	0.887 (0.062)	0.983 (0.021)	0.935 (0.035)
	FPR	0.532 (0.105)	0.552 (0.086)	0.542 (0.092)
	bAUC	0.730 (0.039)	0.774 (0.025)	0.752 (0.019)
Group graphical lasso				
	TPR	0.876 (0.068)	0.984 (0.023)	0.930(0.038)
	FPR	0.536 (0.103)	0.556 (0.087)	0.546 (0.091)
	bAUC	0.758 (0.040)	0.819 (0.019)	0.788 (0.021)
Bayesian edge regression				
	TPR	0.479 (0.095)	0.957 (0.033)	0.718 (0.050)
	FPR	0.047 (0.023)	0.267 (0.045)	0.157 (0.023)
	bAUC	0.803 (0.037)	0.912 (0.020)	0.857 (0.019)

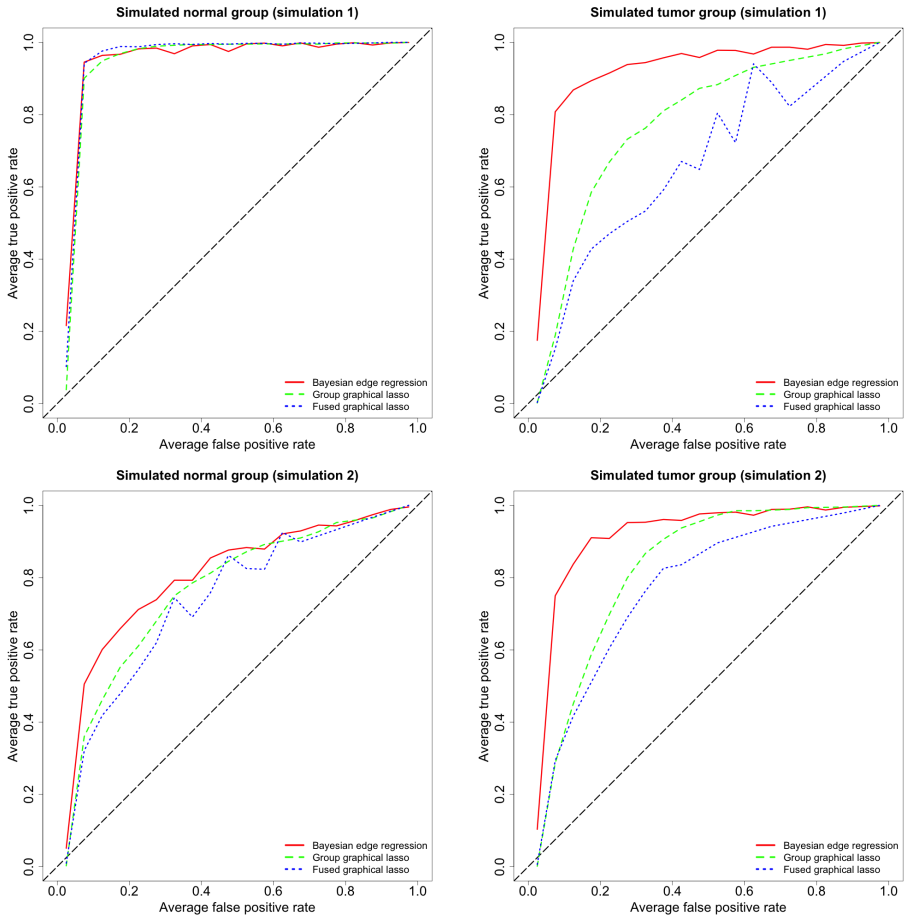


Figure 3.2 : Simulation of Section 3.5.1. ROC curves for structure learning of simulated normal and tumor graphs in *Simulation 1* and *Simulation 2*.

3.5.2 Case II: categorical exogenous covariates

In this simulation, we just consider the categorical exogenous covariates. There is a well-known multiple graphical models problem, where the estimated graphs share some common features. We construct graphs with related dependence structures and also add variables to each of them. We include $p = 20$ nodes and consider the following three precision matrices to simulate data from three groups.

Group1. $\boldsymbol{\Omega_1} = \boldsymbol{\Omega_T}$, which is the precision matrix of tumor component in *Simulation 2* of Case 3.5.1.

Group2. $\boldsymbol{\Omega_2}$, where we remove 5 edges randomly from $\boldsymbol{\Omega_1}$ by substituting these 5 nonzero elements with zero and randomly add 5 edges to $\boldsymbol{\Omega_1}$ by substituting these 5 zero elements with values uniformly sampled from $[-0.6, -0.4] \cup [0.4, 0.6]$.

Group3. $\boldsymbol{\Omega_3}$, where we randomly remove 10 edges in $E_1 \cap E_2$ from $\boldsymbol{\Omega_2}$ by substituting these 10 nonzero elements with zero and randomly add 10 edges in $E \setminus (E_1 \cup E_2)$to $\boldsymbol{\Omega_2}$ by substituting these 10 zero elements with values uniformly sampled from $[-0.6, -0.4] \cup [0.4, 0.6]$. E is the set of all possible edges and E_1, E_2 are sets of edges corresponding to $\boldsymbol{\Omega_1}$ and $\boldsymbol{\Omega_2}$, respectively.

Then $\boldsymbol{\Omega_2}$ and $\boldsymbol{\Omega_3}$ is processed using the same procedure in Section 3.5.1 to make it positive definite. Although this procedure is able to guarantee generated matrix to be positive definite, it can bring weak signals to $\boldsymbol{\Omega_2}$ and $\boldsymbol{\Omega_3}$, which makes the estimation even more difficult. All these three graphs have the same sparsity (37 edges) and every pair of two graphs have overlapping edges.

Given these three matrices, we generate three groups of random samples $\boldsymbol{Y_{(k)}}, k = 1, 2, 3$ of size $N_{(k)} = 100$ from the corresponding normal distribution $\mathcal{N}(0, \boldsymbol{\Omega_k^{-1}})$. With grouped data including 300 observations, we have an exogenous covariate $Z = 1, 2, 3$ which is a categorical variable for each observation. We first dummy code Z using

3 binary variables X_1, X_2 and X_3 to correspond to each group k, i.e., $\boldsymbol{Y_n} \in k \Leftrightarrow X_{k,n} = 1, X_{-k,n} = 0$ for observation n. Then we add the interaction terms to borrow strength between the groups. Since we have 3 groups, the number of interaction term is $C_3^2 + 1 = 4$. Our *conditional precision function* $\omega(\boldsymbol{X})$ can be given as follows:

$$\omega^{ij}(\boldsymbol{X}) = \beta_1^{ij} X_1 + \beta_2^{ij} X_2 + \beta_3^{ij} X_3 + \beta_4^{ij} X_1 X_2 + \beta_5^{ij} X_1 X_3 + \beta_6^{ij} X_2 X_3 + \beta_7^{ij} X_1 X_2 X_3 \quad (3.17)$$

The parameterization in equation 3.17 is useful for borrowing of strength in that for a given edge (i, j), β_1^{ij}, β_2^{ij} and β_3^{ij} indicate the unshared strength in each group, β_4^{ij}, β_5^{ij} and β_6^{ij} indicate the shared strength just between two groups, and β_7^{ij} indicate the shared strength among all these three groups. It demonstrates the advantage of a regression form over discrete covariates for borrowing strength. Here is a table for the dummy coding looks like for observations in each group.

We generate 50 simulated data sets and then implement our edge regression model to

Table 3.3 : Table of dummy coding for multiple graphical models (K = 3)

group	X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	$X_1X_2X_3$
k=1	1	0	0	1	1	0	1
k=2	0	1	0	0	1	1	1
k=3	0	0	1	1	0	1	1

estimate the graph structure. For the purpose of comparison, we still apply the fused graphical lasso and joint graphical lasso following the same procedure in Section 3.5.1. For the hyper-parameter $\{M_s\}_{s=1,\cdots,7}$ in our model, we set $\{M_s\}_{s=1,\cdots,3}$ by calculating $\hat{\Omega}_{MLE}$ for observations in each group, set $\{M_s\}_{s=4,\cdots,6}$ by calculating $\hat{\Omega}_{MLE}$ for observations in each union of every two groups and set M_7 by calculating $\hat{\Omega}_{MLE}$ for all the observations. We still evaluate the performance of structure learning for all the methods in terms of TPR, FPR and bAUC. In Table 3.4, we show the TPR, FPR

and bAUC averaged across 50 simulated data sets for the selection of edges of each group. Figure 3.3 shows the corresponding ROC curves. TPR and FPR of Bayesian edge regression in Table 3.4 is calculated from the results using $\kappa = 0.1$ and $\lambda = 0.1$.

Table 3.4 : Results of edge selection for simulated examples in terms of true positive rate (TPR), false positive rate (FPR) and bivariate area under the curve (bAUC). The numbers given in this table are averaged across 50 simulated sets and the standard deviations are given within the parentheses. The last column provides the average value across three groups.

Method		Group 1	Group 2	Group 3	Overall
Fused graphical lasso					
	TPR	1.000	0.791	0.799	0.863
		(0.000)	(0.095)	(0.087)	(0.054)
	FPR	0.504	0.322	0.335	0.387
		(0.071)	(0.088)	(0.095)	(0.079)
	bAUC	0.891	0.864	0.824	0.860
		(0.013)	(0.020)	(0.024)	(0.016)
Group graphical lasso					
	TPR	1.000	0.775	0.782	0.852
		(0.000)	(0.089)	(0.083)	(0.051)
	FPR	0.538	0.368	0.383	0.429
		(0.078)	(0.096)	(0.108)	(0.089)
	bAUC	0.875	0.828	0.807	0.837
		(0.013)	(0.025)	(0.030)	(0.016)
Bayesian edge regression					
	TPR	1.000	0.591	0.559	0.717
		(0.000)	(0.097)	(0.087)	(0.042)
	FPR	0.168	0.048	0.047	0.087
		(0.033)	(0.023)	(0.019)	(0.015)
	bAUC	0.949	0.853	0.831	0.877
		(0.005)	(0.027)	(0.038)	(0.017)

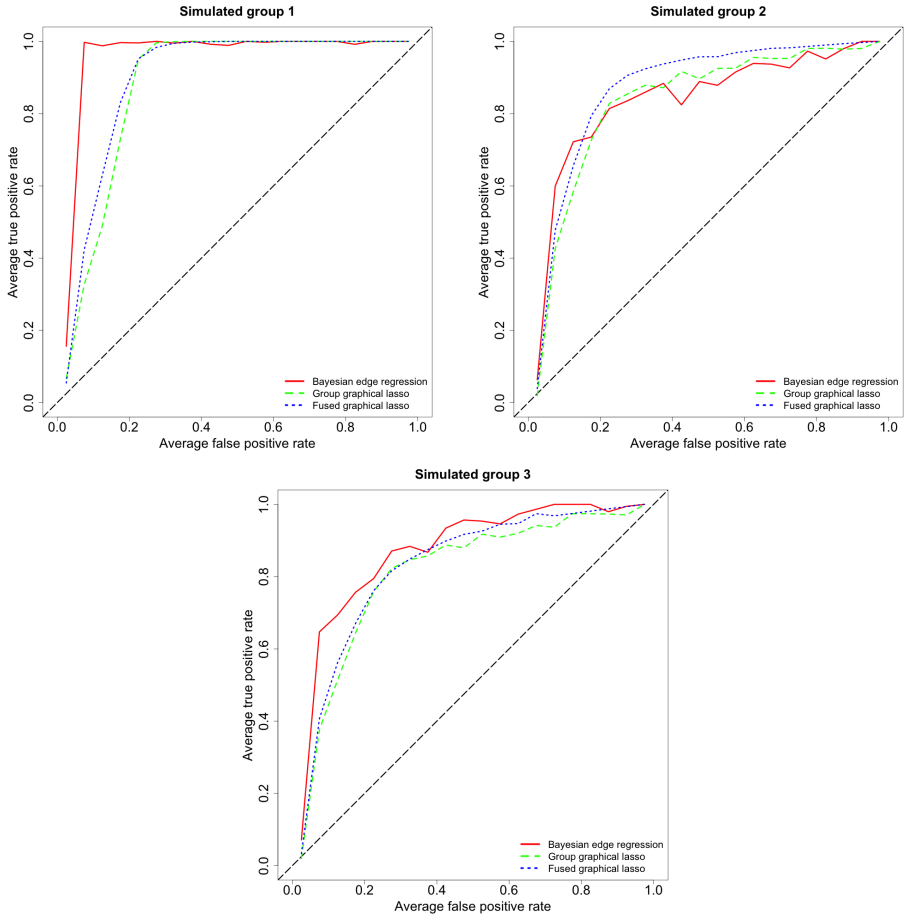


Figure 3.3 : Simulation of Section 3.5.2. ROC curves for structure learning of graphs for simulated three groups.

Results show that both fused and group graphical lassos identify a large number of false edges although they are able to give a high true positive rate. The Bayesian edge regression method has very lower false positive rate but also lower true positive rate. In the setting of multiple Gaussian graphical models, our method is able to improve the specificity at a small cost of sensitivity. According to the measure of bivariate AUC, the Bayesian edge regression method outperforms both graphical lasso methods in terms of edge selection.

3.6 Application

3.6.1 Gene networks of prostate adenocarcinoma given tumor heterogeneity

We performed a case study to demonstrate how to apply our edge regression model in a TCGA cancer data, for which we can respectively construct graphs for normal tissues and tumor tissues to study how the regulatory network varies with the tumor purity in the cancer samples. We downloaded prostate adenocarcinoma (PRAD) data from TCGA data portal (Network et al., 2015b), which consist of RNA-seq gene expression data from normal blood samples and tumor tissues samples. TCGA normal samples are provided as a control group for tumor samples. They should contain no cancerous tissue thus have 0% tumor purity. We choose samples with consistent proportion estimates estimated from different selected gene sets, consisting of 48 normal tissue samples and 211 tumor tissue samples through *DeMixT*. Those estimates are highly consistent with estimated tumor purity from ABSOLUTE, which analyzes somatic DNA alterations (Carter et al., 2012), on those samples (Fig. 3.4). We focus on 87 genes in Androgen Receptor (AR) signaling pathway, which has been found to be a critical determinant of the phenotype of prostate cancer cells (Chen et al., 2008; Network et al., 2015b). In TCGA PRAD samples, the size of normal samples can limit the study of the regulatory networks for the normal tissues, because a number of *covariance selection* methods cannot work or obtain stable estimates when the number of variables (i.e., p) is larger than the number of samples (i.e., n). But this problem can be overcome by our ER method, because the large size of mixed tumor samples, which comprise normal tissues, can borrow strength to the estimation of network structure for the normal samples. With estimated tumor purity, we applied

our graphical regression model on the expression data of those genes.

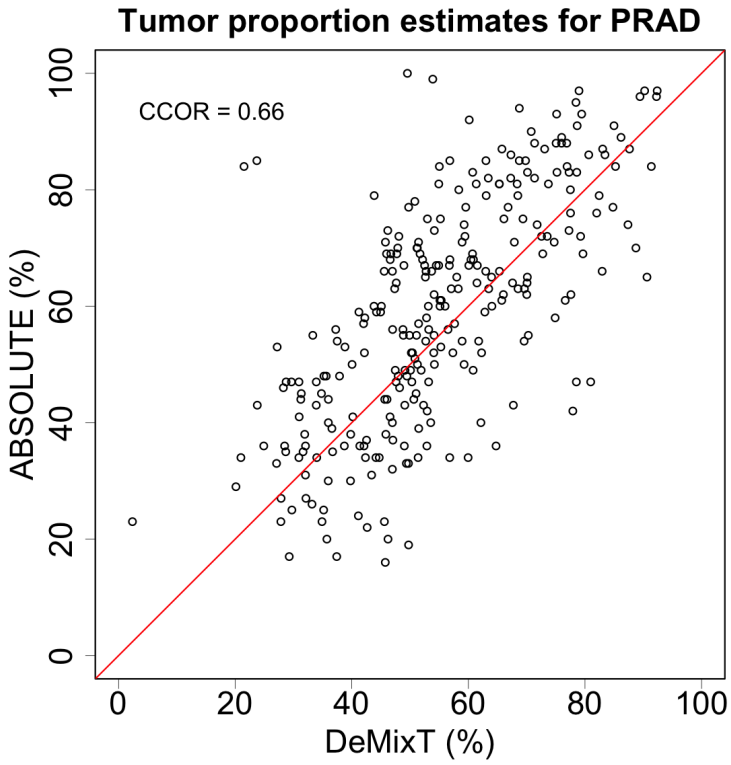


Figure 3.4 : The proportion estimates from *DeMixT* has a high concordance correlation with ABSOLUTE purity estimates, which is inferred from the analysis of somatic DNA alterations.(Carter et al., 2012)

.

We have 259 samples totally with tumor purity ranging from 0 to 1. Pure normal samples are included as a reference group to help the learning of normal structure in tumor samples. Denote the tumor purity with π_n for each sample n. We parameterized our conditional dependence function in the same way as the simulation and run the MCMC sampler according to the Algorithm 2. We ran MCMC samplers for $20,000$ iterations, where the first $10,000$ is a "burn-in" period, the chain was thinned

to every tenth iteration for inference purpose. To obtain enough sparse graphs, we choose $\kappa = 0.15$ and $\alpha = 0.1$. We also follow previous simulation in our simulation to recover the normal graph, the gene regulatory network for healthy tissues, and tumor graph, the network for cancerous tissues, and present the constructed tumor and normal graphs in Figure 3.5. We find that the tumor graph has much more connected edges than normal graph, which suggests AR signaling pathway plays an important role of prostate cancer growth.

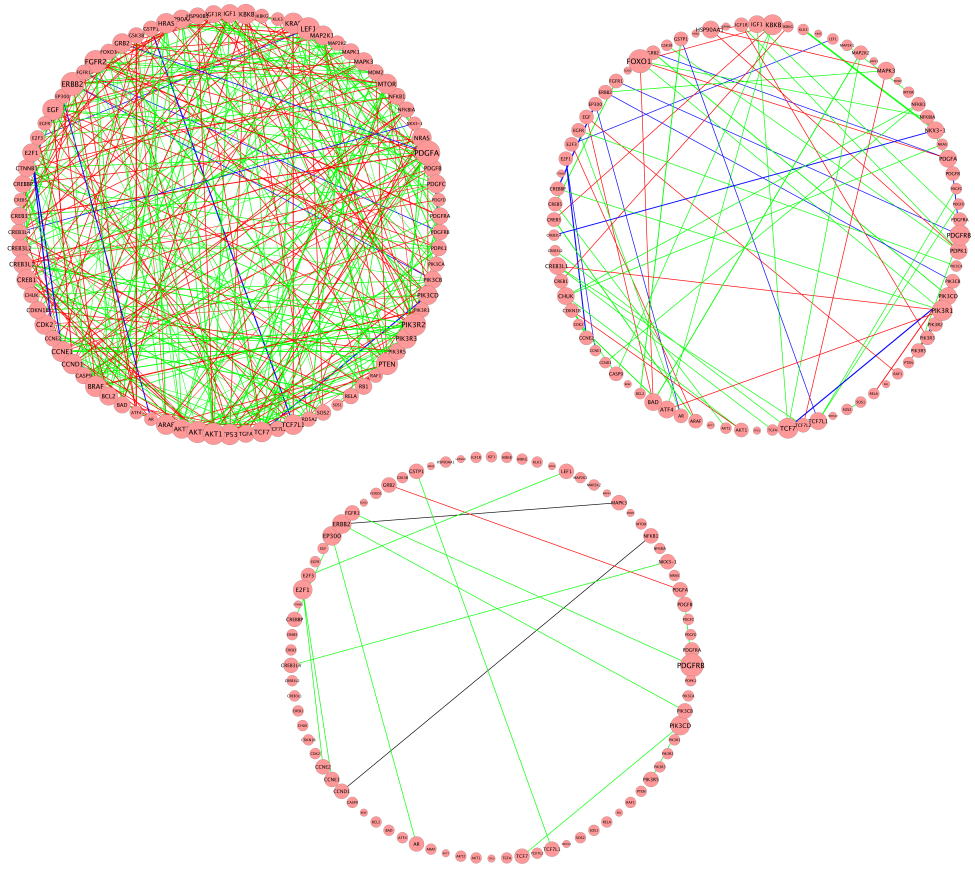


Figure 3.5 : Recovered gene regulatory network of AR signaling pathway in prostate cancer data for cancerous tissue and normal tissue. Positive edges are colored with green and negative edges are colored with red. Common edges of two compared graphs are provided, where edges with consistent sign are colored with blue and different sign with black. Upper left: tumor graph; Upper right: normal graph; Lower: Common edges.

3.6.2 Proteomic networks in hepatocellular carcinoma

In this section, we illustrate the application of edge regression model to the inference of real-world genomic networks, where a set of exogenous covariates plays crucial roles to account for the subject-level diversity of samples. Besides applying edge regression for mixed tumor samples, we also use our model to respectively construct tumor and normal graph without edge regression. We compare all those graphs and investigate

their common structure and differential edges. We ran MCMC samplers for $10,000$ iterations of a "burn-in" period followed by $10,000$ iterations, which were thinned to every tenth iteration for inference purpose. In posterior inference, we used our scheme to select edge and calculate its strength by setting $\kappa = 0.1$ and $\alpha = 0.1$.

Hepatocellular carcinoma (HCC) is a primary liver cancer that causes liver-related mortality and has an increasing incidence in developing countries. Liver cirrhosis, which is commonly caused by hepatitis B and hepatitis C infection, is considered as the primary factor leading to HCC, while HCC is also the major cause of death for patients with compensated cirrhosis. Because cirrhosis is a precursor for HCC, and patients with cirrhosis have high risk for HCC, the study of HCC always involves the consideration of cirrhosis (Fattovich et al., 2004; Sanyal et al., 2010). Recent studies have exposed a number of potential molecular mechanisms that is implicated in HCC. It was reported that inflammation, metabolic process, immune response, growth factor and activation of angiogenesis inherently are associated with the activity in liver cancer progression and development (Dhanasekaran et al., 2016; Aravalli et al., 2008). Cytokines are produced by cells in the liver and released into the blood. For patients with liver disease such as hepatocellular carcinoma, the observed cytokine distribution in the blood may be a mixture of cytokines released by healthy liver cells and released by the diseased liver cells. Once again, by estimating the "tumor purity" and accounting for it in this context we can potentially gain power to detect differences between the cytokine profiles produced by HCC and healthy liver cells. To explore the cytokine distribution in hepatocellular carcinoma patients, we ran the CytokineMAP (Myriad RBM, Austin TX) on plasma samples from 767 HCC patients and 200 normal controls. This CLIA-certified platform is an immunoassay that measures values for 305 Cytokines from inflammation, immune, metabolism, growth hormone, and

angiogenesis pathways. We are interested in comparing the cytokine network emanating from tumor with that emanating from normal liver. For this work, we focus on the network for 10 cytokines in the metabolic pathway.

Cytokines are groups of small proteins released from cells in order to influence the function of other cells. Liver disease such as HCC can be expected to affect the cytokine distribution and networks, and cytokines produced by the liver are secreted into the bloodstream and be detected. Applying deconvolution algorithms to the normal and HCC patients, we found that a sizable number of HCC patients have cytokine distributions like the normal controls, while others are very different. Some HCC patients have cytokine profiles very similar to subjects with normal liver function. Others have highly aberrant cytokine profiles. For each HCC patient sample, we apply our newly developed computational tool *DeMixT* for tumor deconvolution, which we claims to outperform previous methods with respect to estimation accuracy in Chapter 2. From this tool, we can obtain π_i, a measure of heterogeneity representing proportion of the cytokine signal that appears to be from tumor. In the other word, π_i measures the degree of aberration in cytokine profile relative to normal, and is measure of aggressiveness of disease. Patients with $\pi_i = 0$ have cytokine profiles like normal controls, patients with $\pi_i = 1$ have cytokine profiles characterizing HCC that differ the most from normal, while other patients have $\pi_i \in (0, 1)$ that can be interpreted as proportion of cytokine expression in the plasma emanating from tumor versus normal liver graphs we recover in this example can be used to describe HCC induced cytokine distribution. It has been shown that the tumor purity for HCC patient samples (Fig. 3.6) has a distribution with heavy weights around 0 and 1. It suggests that some HCC patients have small aberrations in cytokine, while other have more aberrations towards tumor. This attractive behavior of tumor purity mo-

tivates us to compare cytokine graph in patients with normal liver function versus HCC induced function, focusing on the most highly aberrated. We ran our model to estimate normal graph, the graph of normal cytokine distribution, and tumor graph, the graph of cytokine distribution in HCC patients with most aggressive cancer in our following analysis.

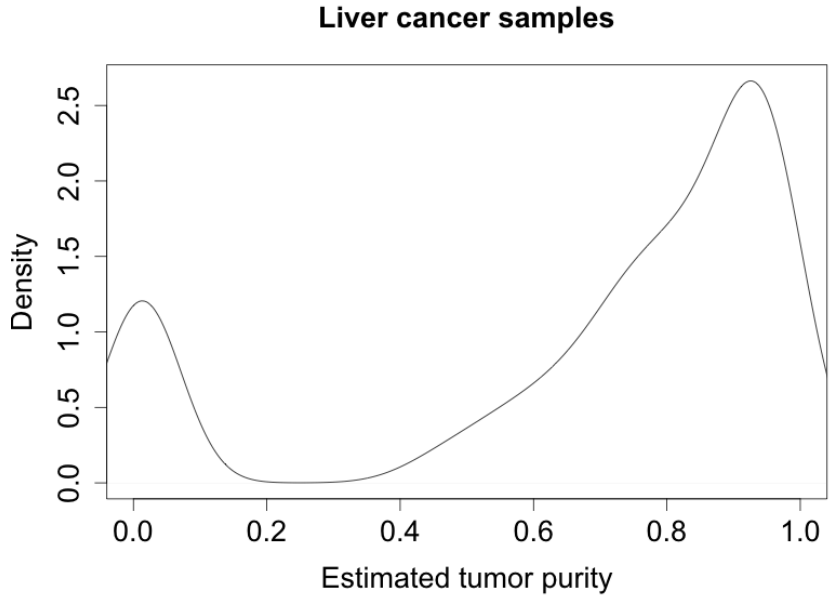


Figure 3.6 : Density plot of estimated tumor purity of HCC tumor samples through *DeMixT*.

If we estimate the cytokine network for normals and HCCs without accounting for this heterogeneity, the tumor network will be attenuated by patients whose cytokine distribution is normal. Thus, we would like to estimate the normal cytokine network and a pure HCC cytokine network that weights patient samples according to π_i. We introduced our edge regression method that can be used to estimate the normal and tumor graphs and test for differential edges.

We applied the same parameterization as in the Section 3.5.2 by setting normal-group specific coefficient α and tumor-group specific coefficient β. Normal samples are included as a reference group for inference of structure of normal graph. We performed the normalization by centering the mean to one and scaling the standard deviation to one and then run our algorithm and make posterior inference. Following Section 3.5.2, we focus on 67 cytokines that union three critical signaling pathways, the growth hormone (GH) signaling pathway, angiogenesis signaling pathway and metabolic signaling pathway, and recovered their normal graph and tumor graph by analyzing 967 HCC samples with their estimated tumor purity. For purpose of comparison, we also performed separate inferences for all those 767 liver cancer samples to get a tumor graph without accounting for the tumor heterogeneity, and for all those 200 normal samples to recover a normal graph. These analyses based on our edge regression model will provide a better understanding of how the tumor purity affect the tumorigenesis of HCC.

We used the same prior setting as in Section 3.5.2, and chose tuning parameter σ_λ that provides acceptance rate of Metropolis step at around $20\% \sim 30\%$. For convergence diagnostics, we first checked the trace plots of all the parameters, and they show good mixing of MCMC chains. In particular, we calculated all the $p-$values of Geweke convergence diagnostic for all the parameters. The histogram of $p-$values under the multiple testing (Fig. 3.7) shows that the $p-$value is uniformly distributed, which indicates our sampling distribution has converged as stationary. After sampling parameters, we obtained the posterior samplers $\{\rho_N^{ij,l}\}_{l=1}^L$ and $\{\rho_T^{ij,l}\}_{l=1}^L$ by calculating through $-\omega^{ij,l}(\pi)/\sqrt{\omega^{ii,l}\omega^{jj,l}}$ given $\pi = 0$ and 1. Then we ran our FDR-controlling procedure to select edges and calculate the posterior means of partial correlations for those selected edges. We chose $\kappa = 0.1$ because it gives a reasonable sparsity in

the resulting graphs. Heatmap plots are provided under different κ for comparing the sparsity of constructed tumor graphs, normal graphs and differential edges in Appendix B. (Fig. B.1, Fig. B.2 and Fig. B.3).

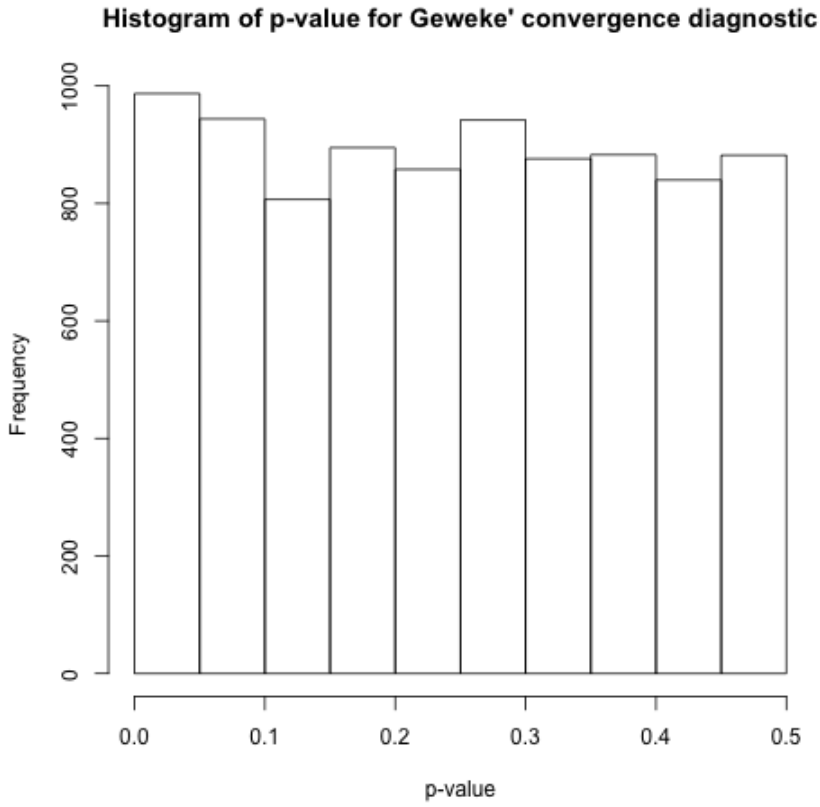


Figure 3.7 : Histogram of $p-$values under Geweke convergence diagnostic for all the parameters we sample from the MCMC chain.

We discover 87 new edges in the inferred networks by applying our ER model, among which 83 edges appear in either tumor or normal graph, 2 edges appear in both tumor and normal graphs but with opposite signs of correlation, and 2 edges appear in both graphs with the same signs of correlation but with obvious difference in scale of partial correlation.

We report several new edges detected through ER that have interesting differences that vary with tumor purity (Fig. 3.8). In each figure, the top half panel presents the linear curve fitted between edge strength calculated from posterior mean and tumor purity, with its 95% credible interval; the bottom half panel shows how the posterior probability of inclusion vary with the tumor purity. The first two of them show that the conditional dependency between CA-15-3 and MCP-1 and that between MIP-1,alpha and MIP-1, beta have different signs in tumor and normal graph, which suggests an opposite regulatory relationship for those cytokines. Regulation relationship between CA-15-3/MCP-1 disappears between $0.21 \sim 0.63$, and that between MIP-1,alpha/MIP-1, beta disappears between $0.1 \sim 0.79$. Some other interesting edges we present for 6Ckine/Resistin, Adiponectin/FSH, IFN-gamma/MIP-1,beta appear on the normal graph but not the tumor graph. We also report that those edges for Decorin/MIP-1,beta, FABP,adipocyte/IGFBP-3, HB-EGF/Kallikrein5, IGFBP-2/IGFBP-3, HGF receptor/VEGFR-1, IL-2/LH appear on the tumor graph but not the normal graph. We also note that IL-2/IL-10, which although have consistent signs, are reported with large-scale difference in the tumor and normal graph. We believe those regulatory relationships are noteworthy for future research.

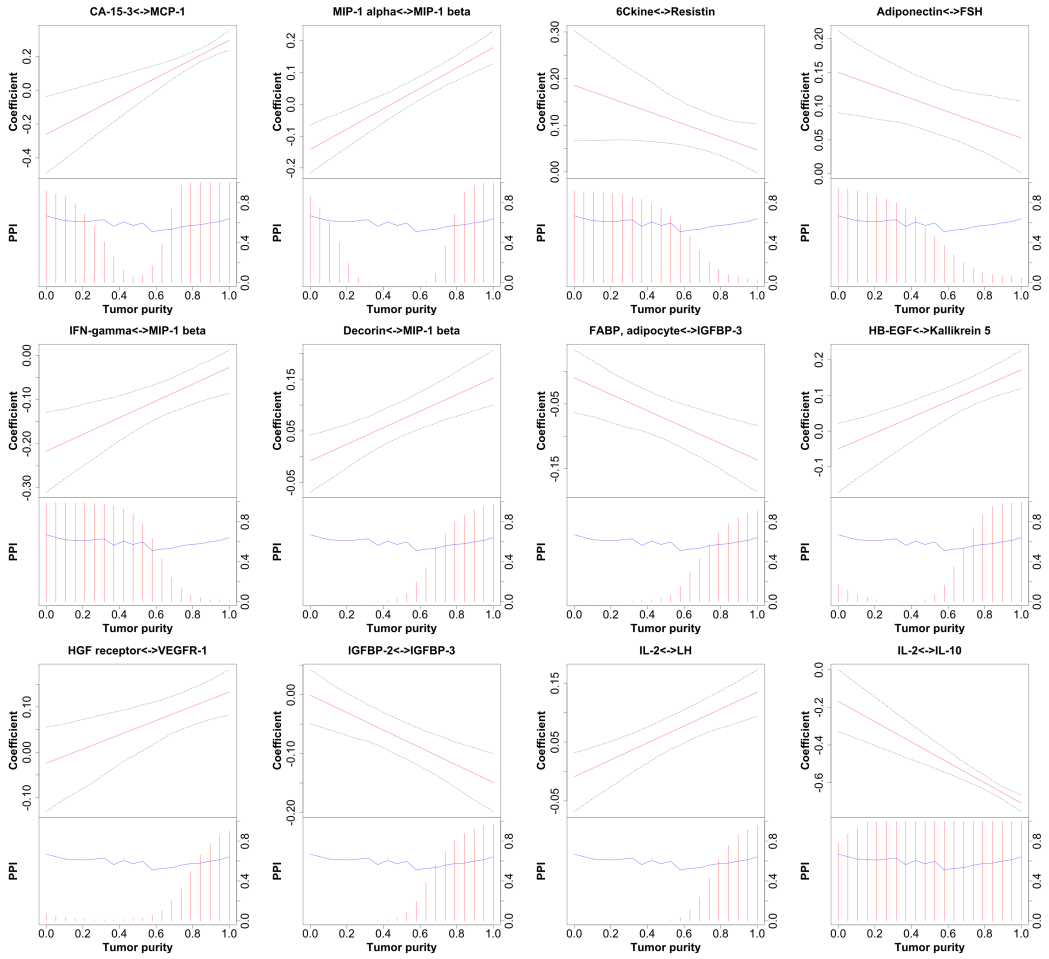


Figure 3.8 : Some new edges detected through ER. The linear curve with 95% credible interval between the edge strength (posterior point estimate of ρ^{ij}) and tumor purity are shown in the top portion. The bottom portion describes how the posterior probability of edge inclusion changes with tumor purity. Blue lines is the probability cutoff given from global FDR controlling procedure at level $\alpha = 0.1$ under different tumor purity.

The resulting graphs are presented in Figure 3.9, while we also include graphs by just running on tumor samples without setting additional covariates in the model (Fig. 3.10). In these figures, the size of edge is proportional to the magnitude of estimated partial correlation between corresponding cytokines, and the size of node describes the degree of connectedness of the corresponding cytokine.

Hepsin is identified as a hub node with a degree of 16 in the tumor graph, while it just has 4 connected edges in the normal graph (Table B.2). Compared with Figure 3.10 , we find that node Hepsin cannot be identified with many edges that are connected between angiogensis pathway and metabolic pathway in the tumor graph, if tumor purity is not given to implement edge regression. It shows that edge regression is able to identify more strong edges. Although tumor graph that is estimated without accounting for tumor heterogeneity shares most edges with that given by edge regression, those identified edges are different regarding sign and strength. Furthermore, the difference of normal graphs suggest that the signaling function in the normal component of tumor samples can also be affected by the tumor microenvironment, which also implies that the study of normal tissues in tumor samples potentially differs from that of normal tissues in healthy samples in cancer research.

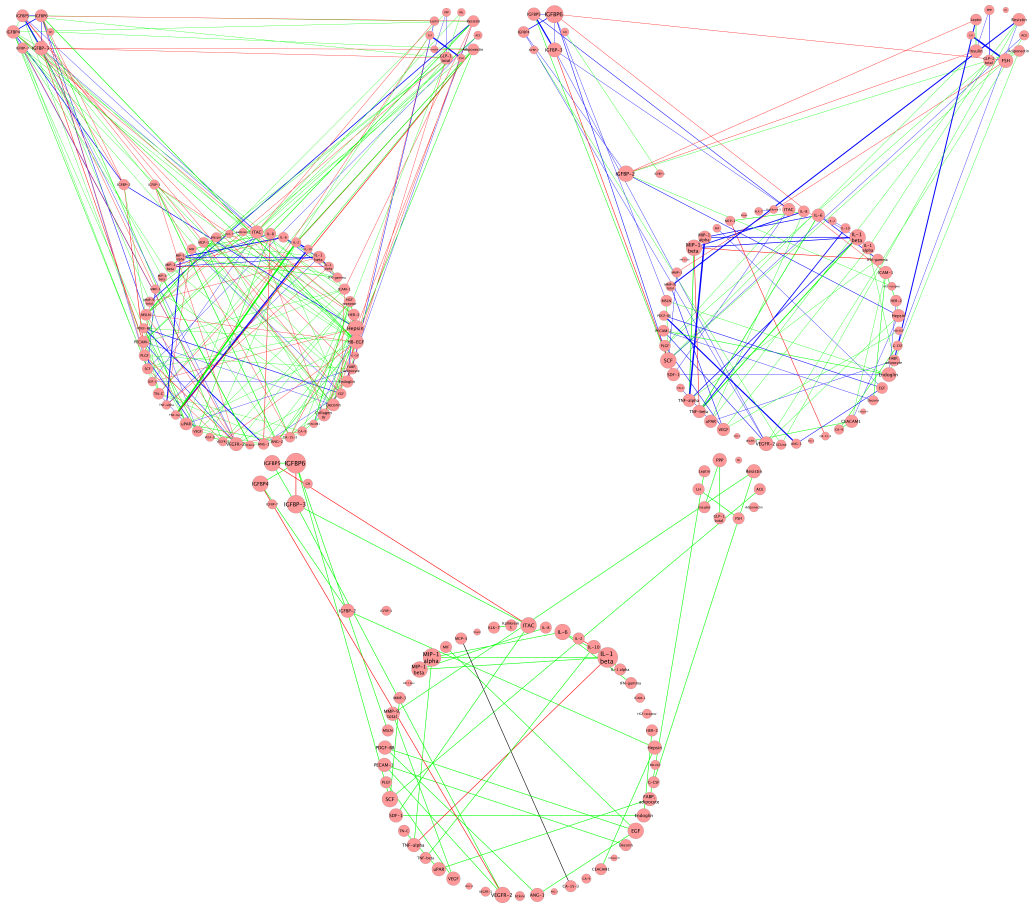


Figure 3.9 : Inferred cytokine signaling pathways-the GH, Angiogenesis and Metabolic pathway through applying ER. Positive edges are colored with green and negative edges are colored with red. Common edges with consistent signs are colored with blue and different signs are with black. Upper-left: tumor graph; Upper-right: normal graph; Bottom: common graph.

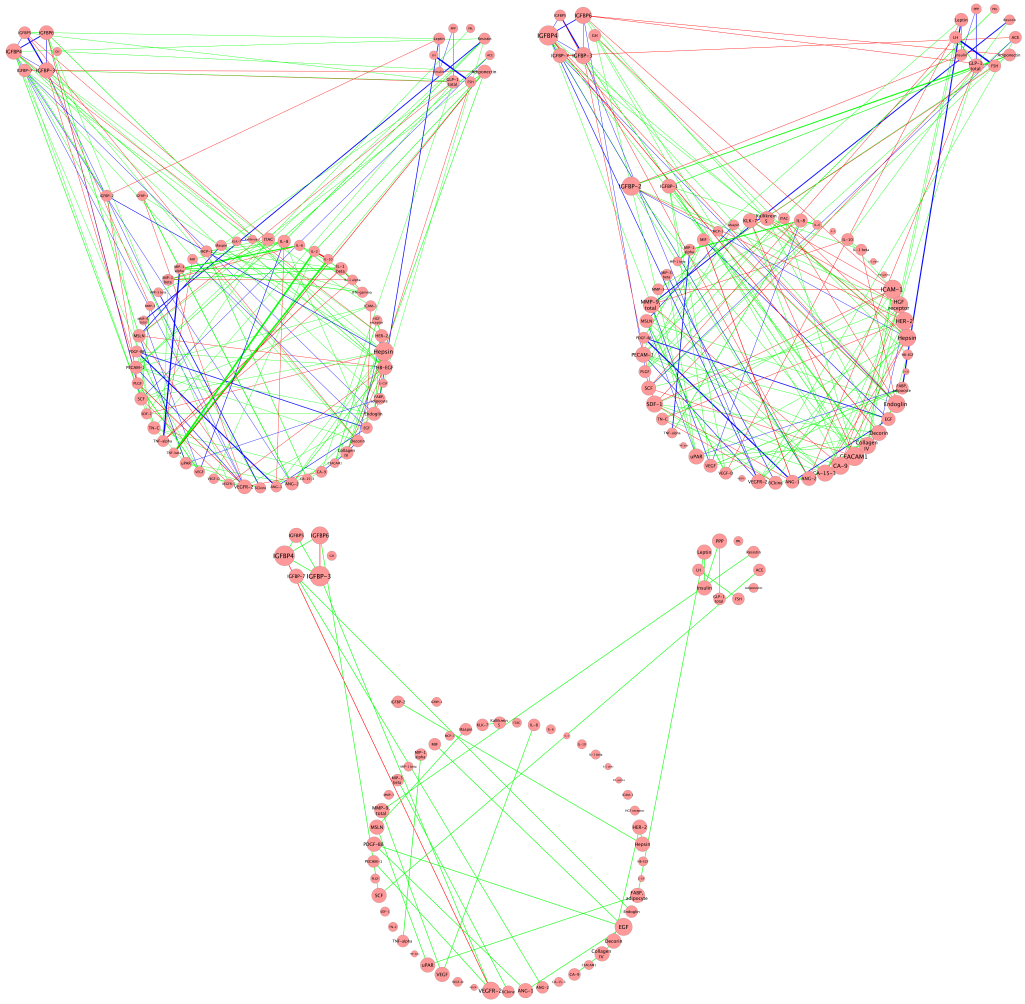


Figure 3.10 : Inferred cytokine signaling pathways-the GH, Angiogenesis and Metabolic pathway without applying ER. Positive edges are colored with green and negative edges are colored with red. Common edges with consistent signs are colored with blue and different signs are with black. Upper-left: tumor graph; Upper-right: normal graph; Bottom: common graph.

In addition to this analysis on a union of those three pathways, we also applied our model to another two important pathways, the inflammation and immuno pathway, to study the networks for these two separate pathways. We ran our MCMC algorithm and implemented the posterior inference in the same way to recover their tumor and normal graphs. All the $p-$values of Geweke convergence diagnostic indi-

cate good mixings of MCMC chains for sampling distributions for these two pathways (Fig. 3.11). We summarize results given in the figures as below (Fig. 3.12, Figure 3.13, Figure 3.14 and Figure 3.15). ER detect more edges for both tumor and normal graph in the inflammatory signaling pathway, but fewer edges for normal graph in the immune signaling pathway. The hub genes on both tumor and normal graphs also change between applying and without applying ER (Table B.3 and Table B.4). MIP-1, beta is identified as a hub gene of the inflammatory pathway in both tumor and normal graph through ER. Those findings may be noteworthy for future clinical studies.

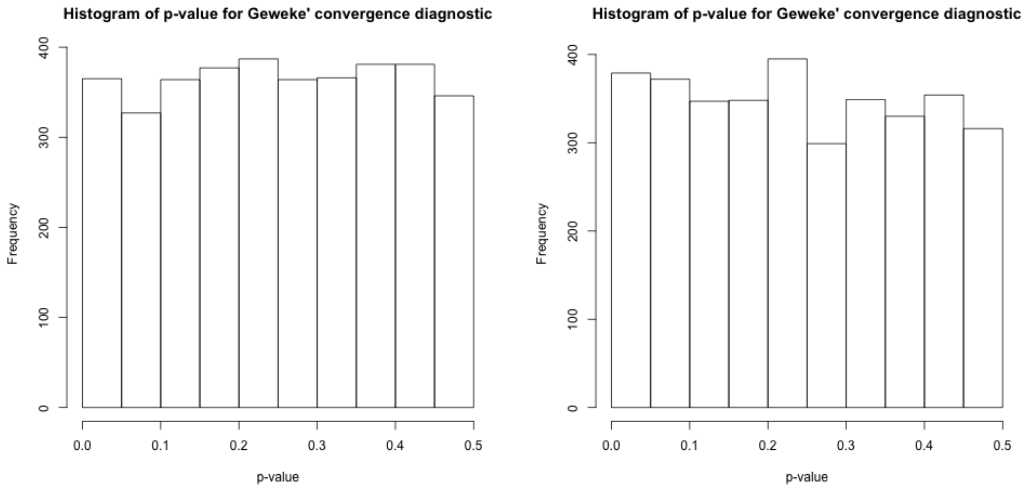


Figure 3.11 : Histogram of $p-$values under Geweke convergence diagnostic for all the parameters we sample from the MCMC chain. The left figure is for the inflammation pathway; the right figure is for the immuno pathway.

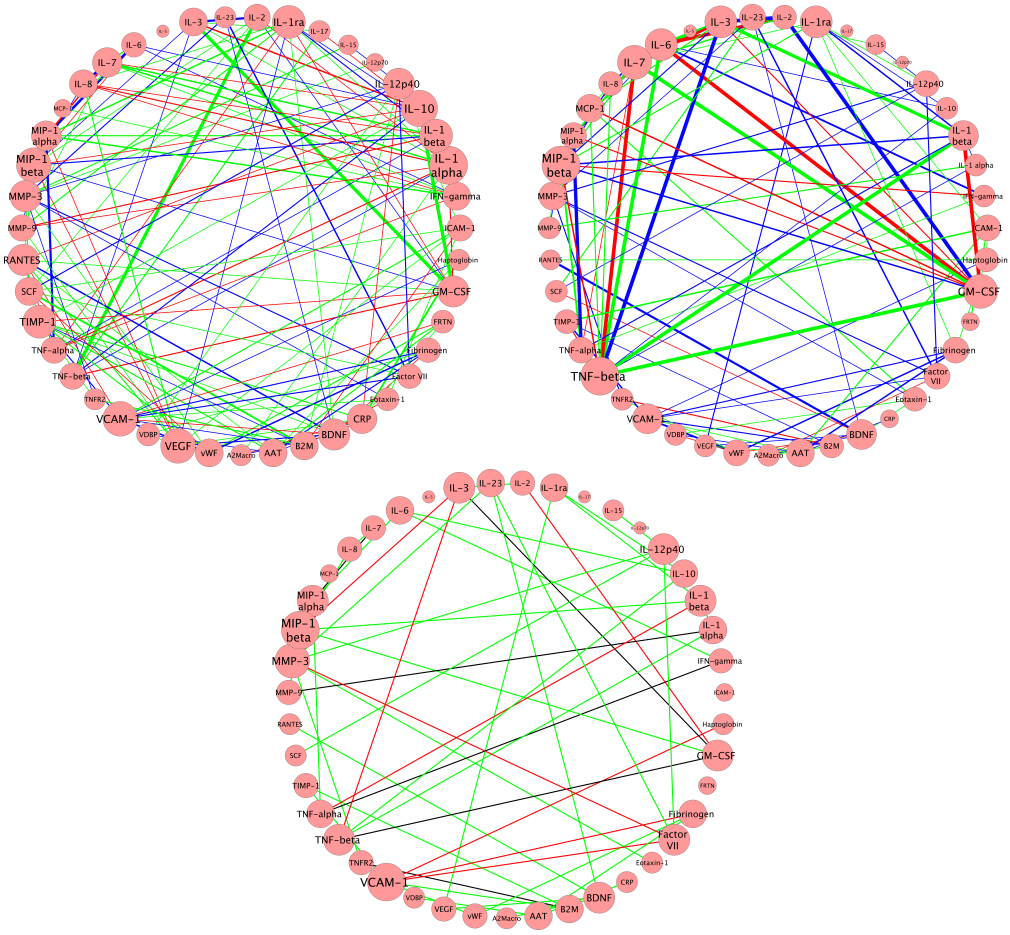


Figure 3.12 : Inferred inflammatory cytokine signaling pathway through applying ER. Positive edges are colored with green and negative edges are colored with red. Common edges with consistent signs are colored with blue and different signs are with black. Upper-left: tumor graph; Upper-right: normal graph; Bottom: common graph.

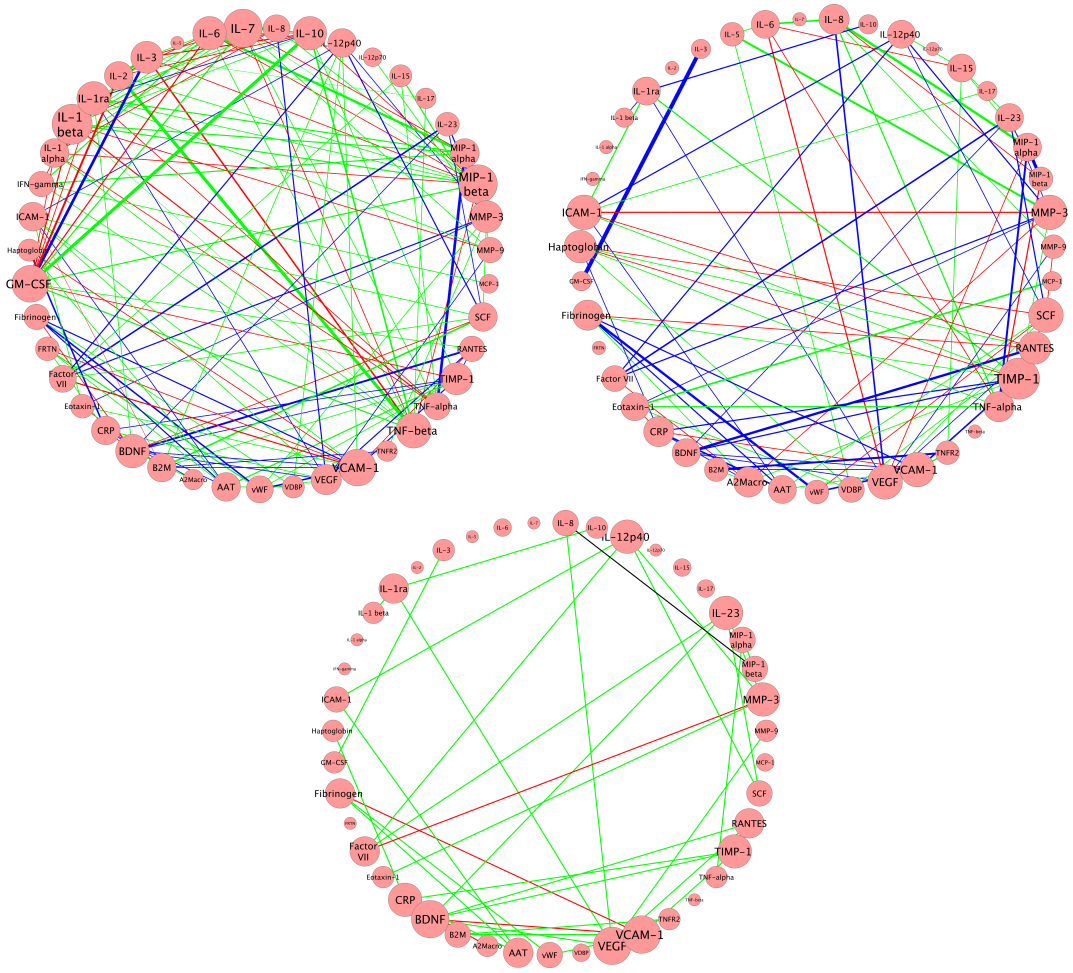


Figure 3.13 : Inferred inflammatory cytokine signaling pathway without applying ER. Positive edges are colored with green and negative edges are colored with red. Common edges with consistent signs are colored with blue and different signs are with black. Upper-left: tumor graph; Upper-right: normal graph; Bottom: common graph.

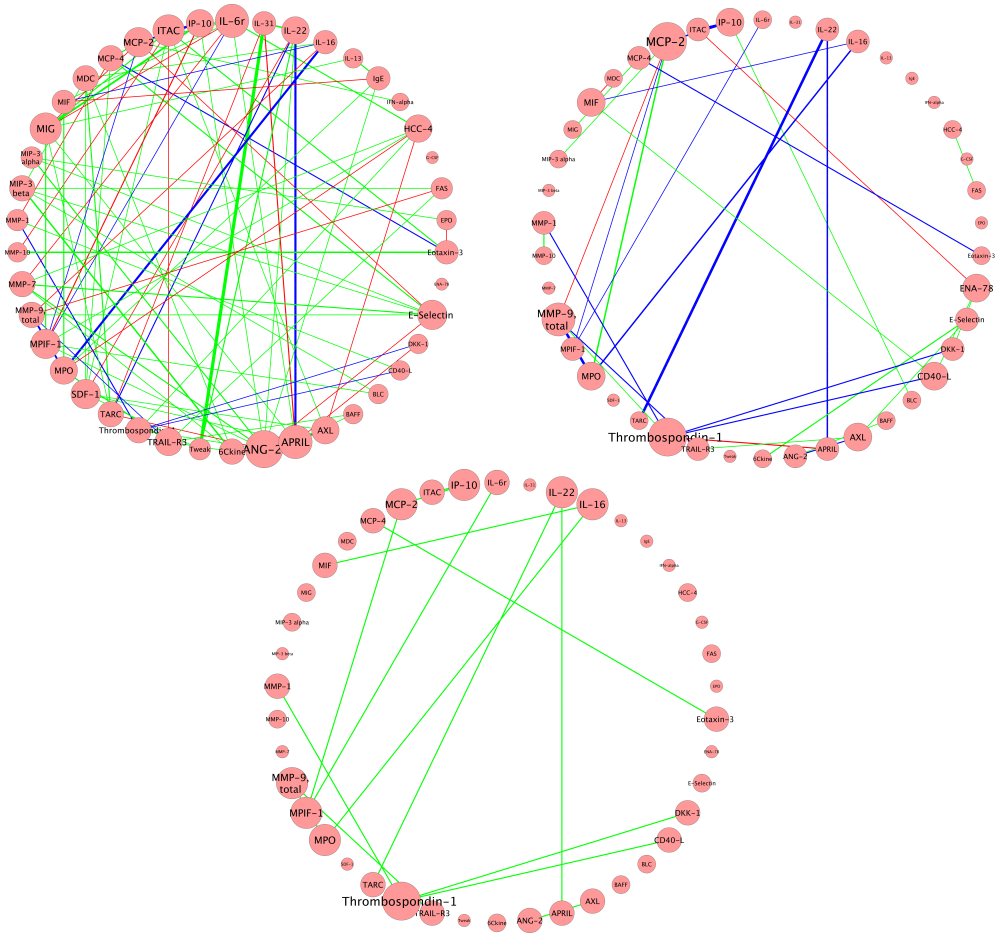


Figure 3.14 : Inferred immune cytokine signaling pathway through applying ER. Positive edges are colored with green and negative edges are colored with red. Common edges with consistent signs are colored with blue and different signs are with black. Upper-left: tumor graph; Upper-right: normal graph; Bottom: common graph.

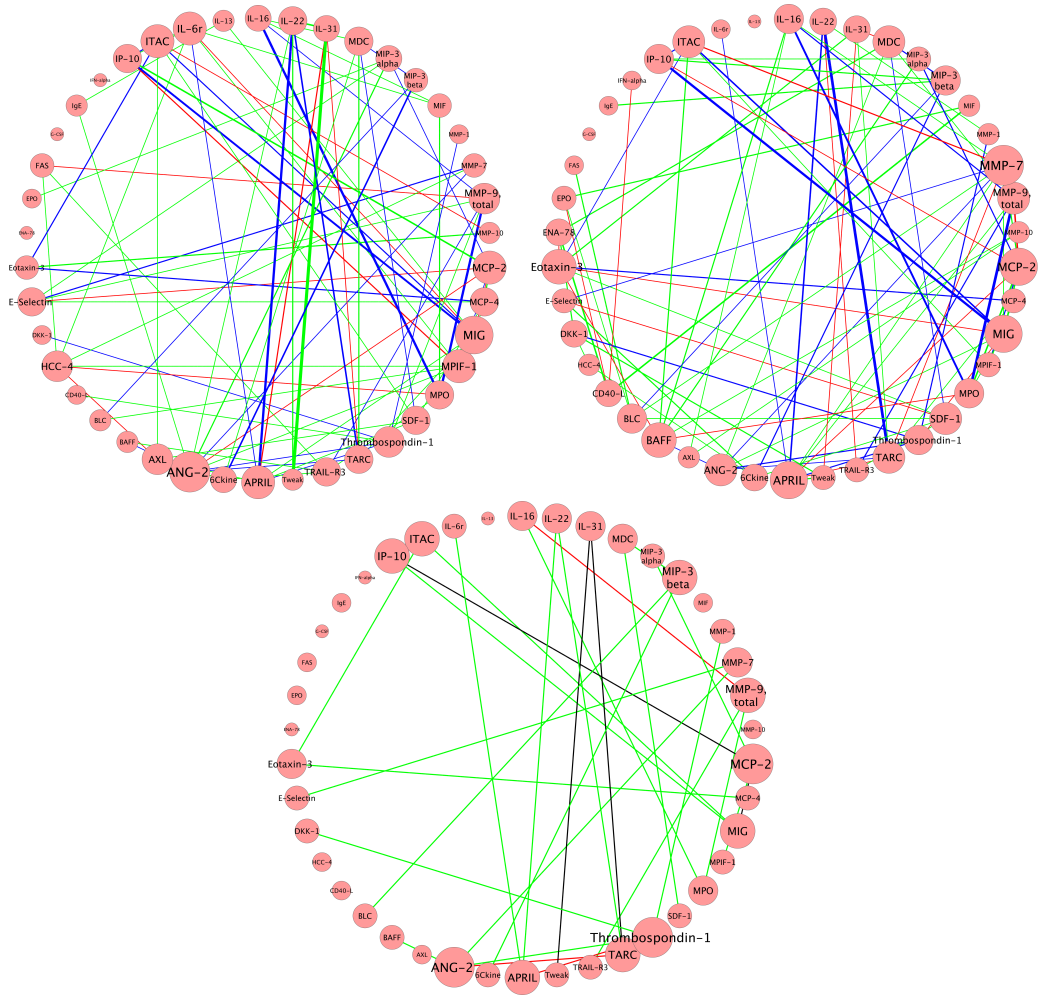


Figure 3.15 : Inferred immune cytokine signaling pathway without applying ER. Positive edges are colored with green and negative edges are colored with red. Common edges with consistent signs are colored with blue and different signs are with black. Upper-left: tumor graph; Upper-right: normal graph; Bottom: common graph.

3.7 Discussions

In this chapter, we introduce a novel model, *Bayesian edge regression*, for construction of non-static undirected graphs varying with exogenous covariates. We claim the significance of developing a new tool to include environmental factors into regulatory networks in genomic analyses, especially for cancer research, the results of

which can be significantly skewed. We bridge edge strength with exogenous covariates through our defined *conditional precision function* and apply a joint regression model for estimation. Our proposed method also explicitly guarantees the symmetry of estimated *precision matrix*. The Bayesian adaptive shrinkage approach imposes the sparsity of both *precision matrix* and the relationship between edges and covariates. Our sampling scheme employs a Gibbs sampler, which leads to rapid mixing of MCMC chain. Our hyper priors for normal-gamma shrinkage prior are set to allow the borrowing of information for regularization parameters of covariate coefficient across different edges. Our sampling procedure is also coherent and suggests simple priors for updating hyper-parameters. We show different parameterizations of our *conditional precision function* with its practicality through simulation study. We also demonstrate that our method is able to provide a reasonable sensitivity and specificity for edge selection. The parameterization is flexible and has been shown to be able to borrow strength in a group-specific setting. We then apply our method for the data of cytokine measurements from blood plasma samples, including normal samples and tumors from hepatocellular carcinoma with tumor purity estimates.

Although in our current work, we do not include a discussion of sampling scheme for nonlinear case, the *conditional precision function* is allowed to be nonlinear, and spline-based semi-parametric methods can be employed for the parameterization. We will extend our work to incorporate the nonlinear relationship in the future. Moreover, the *precision matrix* for Gaussian distribution should be positive definite. In reality, the estimator of our method is not guaranteed to be a positive definite matrix. More generally speaking, all regression-based method does not guarantee the positive definiteness, but it does not affect our primary objective for edge selection instead of estimating a matrix. Finally, we believe our work will help in construction

of networks from heterogeneous genomic data.

Chapter 4

Logistic Regression with Scaling Factor Accounting for Tumor Purity

4.1 Abstract

Diagnostic classification of patient samples is an integral part of analyzing gene expression data in cancer research. Logistic regression is a standard tool for prediction of binary outcomes. In genome-wide association study, the number of genes far exceeds the number of observed subjects. Lasso penalized logistic regression is always added to implement the selection of predictive genes for case-control disease in cancer studies. However, tumor tissues consist of a variety of non-cancerous cells, which have been claimed to contaminate the gene expression patterns of tumors in a lot of literature, as well as cancerous cells. Just as what we discuss about in the first two chapters, the proportion of cancerous cells, the tumor purity, varies widely in tumor samples. Thus, when a clinical outcome mainly associated with cancerous cells is predicted for tumor samples, a direct application of logistic regression neglects different contributions for different samples caused by different tumor purity and is highly sensitive to this deviation of underlying assumptions. We propose a logistic regression model with scaling factor that relates the tumor purity to the uncertainty of observation for improving robustness and providing more accurate estimation. Our model is able to quantify the uncertainty of each sample using the tumor purity for both estimation and prediction. We present strategy for fitting our model in both

settings of logistic regression and penalized logistic regression. We show that our method is able to work well through a set of simulations. Finally, we believe our method will reduce the bias introduced by tumor purity in binary classification on genomic analyses for tumor samples.

4.2 Introduction

The more recent development of gene expression profiling technologies to measure whole-genome mRNA abundance has contributed to the genome-wide association study for identification of disease genes and prognostic prediction for patients (Simon, 2003). The vast amount of gene expression data can help the disease classification in a lot of previous attempts of genomic analyses (Dettling and Bühlmann, 2004). It has been shown that patients from acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) can be classified by using microarray data (Golub et al., 1999). Gene expression profiles have also been used to predict clinical status of some human cancer (West et al., 2001). A wide variety of methods have been proposed for building classification models using genomic data, which include classification tree (Dudoit et al., 2002), support vector machine (Guyon et al., 2002), relevance vector machine (Li et al., 2002), Gaussian Process (Chu et al., 2005), partial least squares (Nguyen and Rocke, 2002) and logistic regression. Our work considers the logistic regression, which has become a prevalent method for binary classification by using gene expression data (Cawley and Talbot, 2006; Liao and Chin, 2007; Sartor et al., 2009; Wu et al., 2009; Zhu and Hastie, 2004). In high-dimensional gene expression data, the number of predictors p always far exceeds the number of samples N, thus lasso penalized logistic regression lends strength to feature selection by adding a $L1$

penalty term. This penalty can be tuned to select a set of small number of discriminatory biomarker genes for predicting case-control disease. The use of penalized logistic regression is able to implement a prognostic prediction at the same time with association mapping of biomarker genes.

Genome-wide association study also has important implications of understanding genetic basis for cancers. It is also a primary task to select predictive biomarker genes and classify clinical status associated with tumors in cancer research (West et al., 2001). However, tumor samples show the cellular heterogeneity due to the intricate microenvironment where tumor grows. The clinically derived tumor samples consist of abundant non-cancerous cells, including a variety of stromal and immune cells. The content of cancerous cells, which is termed as tumor purity, can vary widely among different samples. There have been a few studies to disclose the potentially negative effects of tumor purity in clustering of cancer subtypes (Aran et al., 2015). Tumor purity problem can introduce significant bias into genomic analyses, including the predictive classification of tumor samples. The non-cancerous component in tumor contaminates the expression patterns after profiling, which may under-detect the gene expression signatures associated with the clinical outcomes for cancer prognosis. Implementation of conventional classification method could yield misleading results without taking the tumor purity into account. A number of deconvolution methods for purifying individual gene expression profiles have appeared to aim at this problem, including *DeMixT* we mention in Chapter 2. They are claimed to be able to recover the expression patterns for cancerous component as well as estimate tumor purity, which helps the logistic regression applied to high-dimension genomic data (Nikooienejad et al., 2016). However, we find that our recovered expression profiles also have systematic biases that increases with the decrease of tumor purity

(Fig. 4.1). It makes sense because deconvolution results are artificial, thus more pure tumor samples could have more power for recovery of cancerous signals. Therefore, for whichever expression data we choose to do classification, raw or deconvolved, it is important to build a classification model based on logistic regression that is able to account for tumor purity for quantifying the power of individual observation in regression.

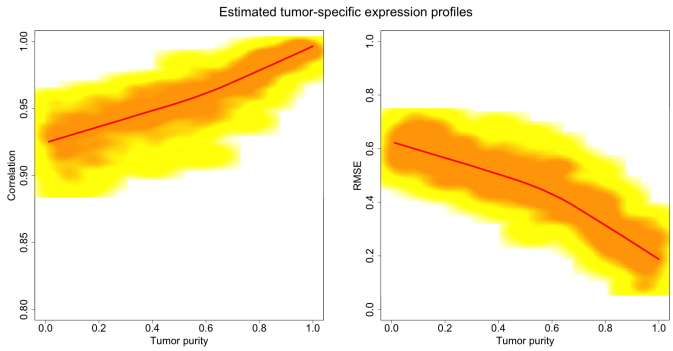


Figure 4.1 : Estimated tumor-specific expression profiles from *DeMixT* present biases that varies with tumor purity. It shows how the Pearson correlation (COR) and root mean square error (RMSE) between estimated expression values and truth for each sample change with tumor purity in a simulation study of deconvolution. Estimation for samples with higher purity are more precise than those with lower purity.

Robust regression is designed to circumvent the limitation of conventional regression method when data are contaminated and model assumption is violated. Weighted least squares (WLS) is a well-known technique to remedy the heteroscedasticity of observed data in linear regression. Under normality WLS estimators can be transformed to maximum weighted likelihood estimators (MWLE), which maximizes a weighted version of log-likelihood function (Vandev and Neykov*, 1998). The weighted methods, with underlying heuristics from MWLE, have also been applied in Bayesian regression model, which is known as power prior (Ibrahim and Chen, 2000). There

have been several robust estimator developed in logistic regression, and a discussion is presented to investigate the robustness of those estimators according to the contamination in method (Carroll and Pederson, 1993). A method of maximum weighted has also been proposed and discussed for logistic regression (Šimecková, 2005). We are motivated by those weighted robustness methods and aim to assign more weight to data with larger tumor purity when the likelihood function is maximized. However, MWLS is unable to explain the contamination in test data for prediction after fitting the regression model, and fails to investigate how the uncertainty affects the regression, therefore wastes the provided data of tumor purity for tumor samples used for prediction. Furthermore, the choice of a proper set of weights may heavily affect the robustness of estimation, and remains open to be discussed for the specific biological problems (e.g. analyze read counts according to the quality of RNA samples) (Law et al., 2014).

In this chapter, we develop a logistic regression model with an inherent scaling factor function instead of directly weighting the likelihood function. This scaling factor helps to quantify the power of data through relating it to the tumor purity for both estimation and prediction. We provide a preferable form of scaling factor function, and then show our strategies for estimating scaling parameter in logistic regression and tuning it in penalized logistic regression through cross validation. Finally, we validate our model by in binary classification through simulation studies.

The remainder of this chapter is organized as follows. In Section 4.3, we give a formal description of logistic regression with its latent variable interpretation. Then we flesh out a scaling factor model with strategies for fitting parameters in logistic regression and penalized logistic regression. We include a set of simulation studies in Section 4.4. Section 4.5 provides the concluding remarks.

4.3 Methods

4.3.1 Logistic regression

In genomic studies of binary outcomes, the response variable Y_i can be coded as 1 for cases and 0 for controls. For each observation i, we have a $p-$dimensional predictor vector $\mathbf{x_i} = (1, x_{i,1}, x_{i,2}, \cdots, x_{i,p-1})^T$, where $x_i \in R^p$ is the gene expression profiles in our discussion, so p can be up to thousands or tens of thousands. Let $X = (\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n})^T$, which is an $n \times p$ matrix. Given a random vector $y_n = (y_1, y_2, \cdots, y_n)$, where $y_i \in \{0, 1\}$, logistic regression is used to model the posterior probabilities of the two classes through a linear expression of $\mathbf{x_i}$.

$$\log \frac{Pr(Y_i = 1|X_i = x)}{Pr(Y_i = 0|X_i = x)} = x_i^T \beta \tag{4.1}$$

where $\beta = (\beta_0, \beta_1, \cdots, \beta_{p-1})^T$ is a $p-$dimensional vector of regression coefficients. The probability $Pr(Y_i = 1|X_i = x)$ of observation i given the predictor vector x_i and coefficient vector β can be written as:

$$p_i = Pr(Y_i = 1|X_i = x) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \tag{4.2}$$

To fit the regression model, we estimate β through maximizing the log-likelihood function

$$L(\beta) = \sum_{i=1}^{n} \{y_i \log p_i + (1 - y_i) \log(1 - p_i)\} = \sum_{i=1}^{n} \{y_i x_i^T \beta - \log(1 + \exp(x_i^T \beta))\} \tag{4.3}$$

We first introduce our scaling factor model for the logistic regression, and then we include a sparse model into our discussion.

4.3.2 Latent variable model

The logistic regression can be interpreted through a latent variable model. Suppose we have a latent variable Y^*. For each observation i, we have

$$Y_i^* = \beta^T X_i + \epsilon_i, \tag{4.4}$$

, where $\epsilon_i \sim Logistic(0, 1)$ is an additive random error variable.

A well-known result can relate the interpretation of logistic regression to a form of ordinary linear regression (Rodríguez, 2007), which claims that the response variable

$$Y_i = \begin{cases} 1 & \text{if} \quad Y_i^* > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

The error variable ϵ_i can be generalized to satisfy a logistic distribution with the scale parameter s, i.e. $\epsilon_i \sim Logistic(0, s)$. Since the logistic distribution is a scale-invariant distribution, we can divide the righthand of the equation 4.4 with the scale parameter s but keep Y^* on the same side of 0. That is, $Y_i^* > 0 \Leftrightarrow \frac{Y_i^*}{s} > 0, \forall s > 0 \Leftrightarrow \beta'^T X_i + \epsilon' > 0$, where $\epsilon_i' = \frac{\epsilon_i}{s} \sim Logistic(0, 1)$ and $\beta' = \frac{\beta}{s}$. Hence, the assumption of generalized logistic distribution for ϵ_i is equivalent to dividing the product of regression coefficients and covariates by the assumed scale parameter s when ϵ_i is assumed to satisfy a standard distribution.

As well known in ordinary linear regression, where $Y_i = \beta^T X_i + \epsilon_i$, the estimate of

weighted least squares, $(X^TWX)^{-1}X^TWY$, which weight the sum of squared residuals or the log-likelihood under the normality assumption for each observation by w_i ($W = diag(w_1, w_2, \cdots, w_i)$), can be used when the errors ϵ_i are assumed under heteroscedasticity, where $\epsilon_i \sim N(0, \sigma_i^2)$, by defining the reciprocal of error variance to be the weights, $w_i = \frac{1}{\sigma_i^2}$. It implies that in a linear regression model, an observation with small error variance should be given a high weight in the fitting process, because it can give relatively more information than an observation with large error variance. Following this idea, given a set of factors $z_i, i = 1, \cdots, n$, which reflect the relative weight of each observation for fitting a logistic regression, we define our scaling factor model for the latent variable in the regression as follows:

$$\begin{aligned} Y_i^* &= \beta^T X_i + \epsilon_i \\ \epsilon_i &\sim Logistic(0, s_i) \end{aligned} \tag{4.6}$$

where $s_i = f(z_i)$ and $f(.)$ is a function to define the relative weight for observation i through z.

It can be equivalently written as:

$$\begin{aligned} Y_i^* &= \frac{\beta^T X_i}{s_i} + \epsilon_i' \\ \epsilon_i' &\sim Logistic(0, 1) \end{aligned} \tag{4.7}$$

Therefore, by phrasing it back to the general logistic function form, we have

$$y_i|\beta, x_i \sim Bernoulli[\frac{\exp(\frac{x_i^T\beta}{s_i})}{1 + \exp(\frac{x_i^T\beta}{s_i})}] \tag{4.8}$$

We can interpret how this model quantifies uncertainty for each sample through relating to additional covariates in terms of the shape of a sigmoid function. Logistic

regression uses a standard logistic sigmoid function to model how the success probability $Pr(y = 1)$ is affected by a set of covariates. A standard logistic function is given by $f(t) = \frac{1}{1+\exp(-t)}$, where in logistic regression $t = x^T\beta$. In our model, the t in the logistic function is scaled by s, which can be interpreted as the steepness of the curve for a more general logistic sigmoid function. Then we have $f(t) = \frac{1}{1+\exp(-\frac{t}{s})}$, where the value of s can affect the shape of $f(t)$. For a logistic regression that uses a classifying cutoff $Pr(y = 1) > 0.5$, a smaller s can stretch this probability of success $Pr(y = 1) > 0.5$ towards 1 and $Pr(y = 1) < 0.5$ towards 0 (Fig. 4.2). Thus, the scaling factor does not affect the binary outcome but affect their "scores" (i.e. the probability of success). Classification score for a sample with smaller s will be adjusted by increasing the positive score and decreasing the negative score (positive: y = 1; negative: y = 0), while both positive and negative score will be flatten towards 0.5 if the sample has comparatively larger s than other samples. It reflects how s controls the power for each sample in estimation of logistic regression. In prediction, s cannot change the final binary outcome, but we can still build it into prediction by "improving" prediction scores after we model s as a function of additional covariates indicating the contamination. Instead of directly weighting log-likelihood using additional covariates for training the model, scaling factor model will intrinsically model the relationship between covariates and the uncertainty in the whole data set.

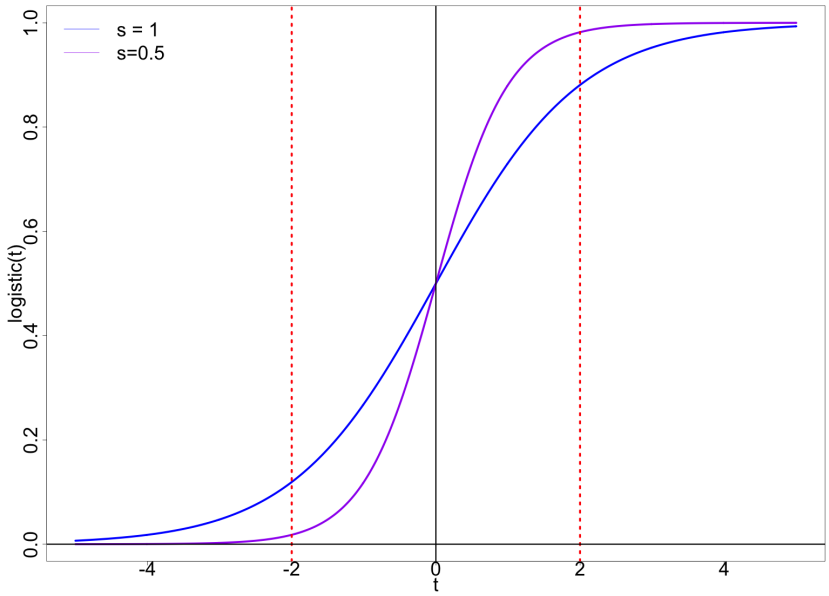


Figure 4.2 : A demonstration plot to illustrate how the scaling factor affect the "score" through the logistic sigmoid function.

4.3.3 Scaling function

As mentioned previously, the function $f(.)$ defines the relative weights of observations through the latent error variance, where small error variance is related to large relative weight. In our problem, since we consider the tumor purity as our scaling factors, large relative weights are desired for observations with larger tumor purity, which indicates more cancerous tissues or more precise deconvolved expression data contributing to cancer diagnosis. Assume we have a set of additional covariates $Z = \{z_1, z_2, \cdots, z_n\}$ that can be used to quantify the contamination in samples. Therefore, $f(.)$ is expected to be a positive non-increasing function of z. Suppose $\frac{s_i}{s_j} = \frac{f(z_i)}{f(z_j)}$ defines a ratio of latent error variance corresponding to a relative weight ratio between observation i and j. We parameterize $f(.)$ for $z \in (0, 1]$ as follows:

$$f(z) = a^z$$
$$0 < a \leq 1 \tag{4.9}$$

Exponential has several properties that meet our demand to define latent error variance through it.

- It is positive for $z \in R$
- It is a decreasing function for $a \in (0, 1)$
- It is degenerated to be equal for all observations when $a = 1$, so weights all the observations equally
- The log-ratio $\log(\frac{f(z_i)}{f(z_j)}) = \log(a) \times (z_i - z_j)$ is linear and the magnitude of a will control the log difference
- It is equivalent to a general exponential function. Given a general exponential function $f(z) = c \times a^{d*z}$, then $\frac{c \times a^{d*z_i}}{c \times a^{d*z_j}} = \frac{a^{dz_i}}{a^{dz_j}} = \frac{a'^{z_i}}{a'^{z_j}}$

We define a as the scaling parameter to calculate a set of scaling factors from z for each sample. In the following sections, we discuss the choice of a for both logistic regression and penalized logistic regression.

4.3.4 An optimization procedure for logistic regression

With our well-parameterized scaling function, we formulate the full model as follows:

$$y_i|\beta, a, b, x_i, z_i \sim Bernoulli[\frac{\exp(\frac{x_i^T\beta}{a^{z_i}})}{1+\exp(\frac{x_i^T\beta}{a^{z_i}})}] \qquad (4.10)$$
$$0 < a \leq 1$$

Although our work mainly aims at penalized logistic regression with scaling function, we first include the estimation for logistic regression with scaling function but without penalty term into discussion. For a that has very small magnitude, it is possible for $\frac{\eta_i}{a^{z_i}}$, where $\eta_i = x_i^T\beta$, to have an intolerable large magnitude for all i. This scaling problem can disconverge the fitting of parameters in logistic regression. We provide two different strategies to bypass it for logistic regression and penalized logistic regression problem. We generalize the scaling function to $f(z) = a^z + b$, where $b \geq 0$ is a shift term to affect the magnitude of $f(z)$. It can slightly affect the ratio of weights between different observations but not affect the other properties mentioned above. With parameters a and b in the scaling function, we have our log-likelihood function as:

$$L(\beta, a, b) = \sum_{i=1}^{n}\{y_i\frac{x_i^T\beta}{a^{z_i}+b} - \log(1+\exp(\frac{x_i^T\beta}{a^{z_i}+b}))\} \qquad (4.11)$$

This objective function is not strictly convex after adding a and b, but we can still design an optimization procedure to find a local optimum for β, a, b and finally compare this solution to $\hat{\beta}'$. $\hat{\beta}'$ represent the MLE estimates of regression coefficients when we fix $a = 1$ and $b = 0$, which is also the original solution without introducing the weights. When we have a best solution of the original logistic regression problem, we only need to see if our current solution improves the estimation performance. We de-

sign an optimization procedure via coordinate descent. We keep the Newton-Raphson step for updating β (Dobson and Barnett, 2008; Venables and Ripley, 2013; R Core Team, 2016), but just insert a step to update a and b through a golden section search with parabolic interpolation. Our optimization procedure is given as follows:

1. Initialize $\beta^{(0)}$, $a^{(0)}$ and $b^{(0)}$

2. Update β, a, b through coordinate descent

   In iteration t

   - Update $x_{s,i}^{(t)} = \frac{x_i}{s_i^{(t)}}$, where $s_i^{(t)} = (a^{(t-1)})^{z_i} + b^{(t-1)}$
   - Update $\beta^{(t)} = ({X_s^{(t)}}^T W X_s^{(t)})^{-1} {X_s^{(t)}}^T W z$,
     where $z = X_s\beta^{(t-1)} + W^{-1}(y - p(x_{s,i}^{(t)}; \beta^{(t-1)}))$, $W = diag\{p(x_{s,i}^{(t)}; \beta^{(t-1)})(1 - p(x_{s,i}^{(t)}; \beta^{(t-1)})\}$, $p(x_{s,i}^{(t)}; \beta^{(t-1)}) = \frac{\exp({\beta^{(t-1)}}^T x_{s,i}^{(t)})}{1+\exp({\beta^{(t-1)}}^T x_{s,i}^{(t)})}$
   - Search $0 < a^{(t)} \leq 1$ and $b^{(t)} \geq 0$ such that $-L(a, b; \beta^{(t)})$ is minimized
   - Until convergence condition is satisfied

Notice in the Newton-Raphson step, we can use a Newton-downhill method to guarantee the objective function $-L(\beta, a, b)$ to be strictly non-increasing in this step. As long as the algorithm is convergent, we keep the estimated a, b and β for the purpose of prediction.

4.3.5 Penalized logistic regression with $L1$ penalty

To implement a feature selection for logistic regression, we subtract a $L1$ penalty term from the log-likelihood function. In lasso penalized logistic regression, we maximize

the objective function w.r.t:

$$\frac{1}{N}\sum_{i=1}^{n}\{y_i x_i^T\beta - \log(1+\exp(x_i^T\beta))\} - \lambda|\beta| \tag{4.12}$$

, where λ is the tuning parameter to control the amount of shrinkage.

Including our scaling function to the penalized logistic regression, we have our objective function to be:

$$\frac{1}{N}\sum_{i=1}^{n}\{\frac{y_i x_i^T\beta}{a^{z_i}} - \log(1+\exp(\frac{x_i^T\beta}{a^{z_i}}))\} - \lambda|\beta| \tag{4.13}$$

Because a can affect the magnitude of β, $L1$ norm of which is bounded by a pre-specified parameters $t = t(\lambda)$, we cannot directly estimate a by maximizing the objective function given the tuning parameter λ. Actually when we estimate a, in each iterative step s_i and β tend to become smaller, which finally fails to converge. The value of scaling parameter may be data dependent. There has been one approach to selecting weights for the WLME through cross-validation (Wang et al., 2005). Similar to its analogy, we also propose an implementation of cross validation to select a as a tuning parameter as well as λ. To bypass the potential convergence problem caused by abnormal scale of $\frac{x_i^T\beta}{s_i}$ for the logistic function, we divide each scaling factor with a normalization constant a^{a_0}. Then we have:

$$\frac{1}{N}\sum_{i=1}^{n}\{\frac{y_i x_i^T\beta}{a^{z_i-a_0}} - \log(1+\exp(\frac{x_i^T\beta}{a^{z_i-a_0}}))\} - \lambda|\beta| \tag{4.14}$$

The parameter a_0 does not affect the relative weights $\frac{s_i}{s_j}$ between two observations, because a_0 will be cancelled off in $\frac{a^{z_i-a_0}}{a^{z_j-a_0}}$. In addition, the scaling function $f(z) = a^{z-a_0}$ still satisfy all these properties we claim above. The choice of a_0 is subjective, and we recommend to use $a_0 = median(z)$ if z is more uniformly distributed. By setting this

in the equation, we have no more than 50% observations with latent error variances no less than 1 and no more than 50% observations with latent error variances no larger than 1. A grid search can then be implemented to select λ and a at the same time following the minimum validation error rule or one standard error rule. In fitting the regression model, we first obtain $\mathbf{x_i'}$ by dividing the predictor vector $\mathbf{x_i} = (1, x_{i,1}, x_{i,2}, \cdots, x_{i,p-1})^T$ by $a^{z_j - a_0}$ for each i (Note: we need to include the intercept term to be scaled by $a^{z_j - a_0}$). The method of cyclical coordinate descent (Friedman et al., 2010) and proximal gradient method (Rockafellar, 2015) are commonly used to find the regularized path for logistic regression. With $\mathbf{x_i'}$ as the predictor vector, in our example we apply the proximal gradient method for regularization. The selection procedure of a amounts to a sequence of nested loops given in Algorithm 2. This cross validation scheme performs a grid search for a and λ.

Algorithm 2 Cross validation scheme for selecting a with λ in penalized logistic regression

H

for increment $a \in (0, 1)$ **do**
    a. compute a sequence of λ_a at which the largest value makes all coefficients shrink to zero given $s_i = a^{z_i - a_0}$.
    b. cross validation
    for increment λ_a **do**
        run the proximal descent method on the penalized logistic problem by scaling the rows of predictor matrix by $(s_1, s_2, \cdots, s_n)$
        compute and record validation error $CV(\lambda, a)$ and standard error $SE(\lambda, a)$ over all folds of validation set
    end for
end for
choose (λ, a) s.t

1. $\hat{\lambda}, \hat{a}$ minimize $CV(\lambda, a)$ **or**
2. one standard error rule: the most sparse model with $CV(\lambda, a) \leq CV(\hat{\lambda}, \hat{a}) + SE(\hat{\lambda}, \hat{a})$

4.4 Simulations

In this section, we include three simulation studies for logistic regression respectively with and without an $L1$ penalty. We compare the performance of our method with the logistic regression model without scaling factor function. Our simulations are designed to correspond to the tumor heterogeneity problem we emphasize in our work, where the datasets are simulated following the model assumption from the previous tumor deconvolution works.

4.4.1 Simulation 1: logistic regression

In this simulation, we include 200 samples, where 150 samples are used as the training data for model fitting and 50 samples are used as the test data for reporting the prediction errors. The element of the true design matrix X is generated from a normal distribution, i.e. $x_{i,p} \sim N(\mu_x, 1)$, where $\mu_x \sim N(7, 1.5)$. We generate an observed design matrix in a way following Chapter 2, which assumes the observed expression profiles from clinically derived tumor samples to be a linear mixture of profiles contributed by pure normal and pure tumor components before log2-transformation of gene expression data. We generate $n_{i,p} \sim N(\mu_n, 0.5)$, where $\mu_n \sim N(7, 1.5)$. We log2-transform a weighted sum of $2^{x_{i,p}}$ and $2^{n_{i,p}}$ with weights π_i and $1 - \pi_i$ to obtain a observed $x'_{i,p}$, which simulates the contamination of $x_{i,p}$ by $n_{i,p}$. The equation underlying simulation is given as below:

$$2^{X'_i} = (1 - \pi_i)2^{N_i} + \pi_i 2^{X_i} \tag{4.15}$$

, where $\boldsymbol{N_i} \sim \mathcal{N}(\boldsymbol{\mu_N}, \boldsymbol{\Omega_N^{-1}})$ and $\boldsymbol{X_i} \sim \mathcal{N}(\boldsymbol{\mu_X}, \boldsymbol{\Omega_X^{-1}})$. $\pi_i \in [0, 1]$ is the purity of X for the observation i.

We assign $\pi = \{\pi_1, \pi_2, \cdots, \pi_n\}$ with an arithmetic sequence from 0.05 to 0.99 respectively in the training data and test data. We set the intercept β_0 to be 1. The following combinations of β are used to generate non-zero coefficient.

$(1)\beta = \{1, 2, 3\}$; $(2)\beta = \{2, 3, 4\}$; $(3)\beta = \{1, 2, 3, 4, 5\}$; $(4)\beta = \{2, 3, 4, 5, 6\}$.

We first standardize the columns in X to satisfy a standard normal distribution. Then we generate $y_i \sim Bernoulli(\frac{\exp(x_i^T\beta+\beta_0)}{1+\exp(x_i^T\beta+\beta_0)})$. We normalize the observed data X' in the training and test data. We run 200 simulations, where each non-zero coefficient setting above is simulated for 50 times.

We compute the average area under the curve (AUC) of the receiver operating characteristic (ROC) curve and the mean squared error of prediction (MSE). MSE is defined by:

$$MSE = \frac{||\hat{y} - y||_2^2}{n} = \frac{\sum_i^n (\hat{y}_i - y_i)^2}{n} \tag{4.16}$$

Table 4.1 summarizes the performance of our logistic regression model with scaling function (LGSF) on mixed data and deconvolved data compared with the original logistic regression model (LG). Results show that LGSF can improve the prediction performance for both mixed data and deconvolved data in terms of AUC and MSE. The logistic regression using deconvolved data is also preferred to using mixed data.

Table 4.1 : Results of binary classification of test set in *Simulation 1* in terms of MSE and AUC with standard errors (SE) in the bracket over 200 simulated datasets.

	Mixed data (LG)	Mixed data (LGSF)	deconvolved data (LG)	deconvolved data (LGSF)
AUC	0.808 (0.007)	0.827 (0.007)	0.810(0.007)	0.838 (0.006)
MSE	0.1770 (0.0015)	0.1707 (0.0014)	0.1760 (0.0015)	0.1655 (0.0015)

4.4.2 Simulation 2: penalized logistic regression

In this simulation, we have a similar setting for generating observed and hidden predictor matrix in the logistic regression. We include $p = 1000$ covariates and have p' covariates with non-zero coefficients, where $p' \in \{3, 5\}$. The four combinations of β in the first simulation are kept in this simulation, and the mixing manner of N_i and X_i for generating X_i' is same. The only difference is that we generate some mixed samples with purity $\pi = 1$, which is considered as the pure tumor samples to stabilize the estimation. After simulating the observed predictor matrix, we also apply our tumor deconvolution tool *DeMixT* to resolve X_i for each sample. We notate the recovered predictor matrix through deconvolution as $\hat{X}_i$. Next, we implement $L1$ penalized logistic regression on X_i, X_i' and $\hat{X}_i$, and apply our scaling factor model on X_i' and $\hat{X}_i$. We assign 200 samples in training data, and 100 samples in test data, and finally compare their prediction performance in the test set. The results are summarized in Table 4.2 in terms of AUC and MSE for both mixed data and deconvolved data using LG and LGSF.

Table 4.2 : Results of binary classification of test set in *Simulation 2* in terms of MSE and AUC with standard errors (SE) in the bracket over 200 simulated datasets.

	Mixed data (LG)	Mixed data (LGSF)	deconvolved data (LG)	deconvolved data (LGSF)
AUC	0.780 (0.007)	0.789 (0.007)	0.795 (0.007)	0.806 (0.007)
MSE	0.206 (0.001)	0.205 (0.001)	0.204 (0.001)	0.198 (0.001)

The large number of features to be estimated in the penalized logistic regression makes prediction difficult to be improved for a contaminated data set. But our model can still improve AUC and reduce MSE for both mixed data and deconvolved data. By using deconvolved data, it can also improve the prediction performance. Finally, we can realize an improvement of 0.026 in AUC through applying our model on the deconvolved data.

4.5 Discussion

In this chapter, we demonstrate that the traditional logistic regression model is not robust to model cancer-associated clinical status without accounting for cellular heterogeneity of tumors. We construct a scaling factor model in logistic regression to incorporate tumor purity for quantifying the uncertainty caused by non-cancerous cells' contamination or the measurement error from deconvolution tools. We also present the strategies to estimate or tune up scaling parameter for logistic regression and penalized logistic regression. Results on the data simulated from the tumor convolution model clearly demonstrate that our proposed model can improve prediction performance compared with traditional logistic regression model. Although our scal-

ing factor model is proposed for logistic regression, it can still be extended to any sigmoid function based model. Furthermore, we propose our model for the problem of tumor cellular heterogeneity, but our model is more general to be applied to problems of classifying contaminated data. For tuning scaling parameters in cross validation, we currently implement a search on a fixed grid of evenly spaced value from 0 to 1. Further work is needed to provide narrower search range for improving efficiency, as well as study the theoretical properties of our model and other modifications of the scaling factor function.

Chapter 5

Conclusions and Perspectives

5.1 Conclusions

The recent advancement of statistical methods proposed for genomic studies, such as classification and correlation, provide deeper insights into the association among the activities of genes under different biological conditions. Cancer researchers benefit from using statistical technique to help them better detect the expression patterns of genes. However, the cellular heterogeneity of tumors has posed a great challenge to the application of conventional genomic methods that analyze gene expression patterns for a homogeneous population. There have been several studies that disclose their significant influences on caner biology by confounding the biological interpretation of genomic analysis results for tumors. One research has shown that acquired resistance to VEGF inhibitor bevacizumab is associated with these gene expression changes that occur predominantly in stromal but not tumor cells, which suggests that understanding stromal signaling pathway is also critical for cancer biomarker study (Cascone et al., 2011). The other study has questioned the potential joint role of JAK3 and CSF1R in a caner-driving pathway after finding that their tandem expressions in bladder carcinoma differed seriously with the tumor purity. They also uncovered falsity on the relative expressions for two important proteins in cancer immunotherapy, CTLA-4 and CD86, between the results taking tumor purity into account and those that do not (Aran et al., 2015). Following those demands for more

powerful statistical methodologies in cancer biology, we put forward a series of topics on statistical modeling for cellular heterogeneity problems. We set out to develop an *in silico* deconvolution method for expression data, which outperforms previous methods in estimation accuracy and account for more variability in normal cell populations. We test our deconvolution method from biologically simulated data and real data, and discover some interesting biological finding through our method. Next, we are motivated to design a new statistical model to serve for the construction of gene regulatory networks of tumor samples. We develop a novel class of undirected graphical models, the edge regression model, which can utilize the tumor purity we obtain from the deconvolution as an indicator of the degree of genomic aberration from normals to investigate the dynamic structure change of networks in heterogeneous tumors. Our edge regression method is firstly proposed for the cellular heterogeneity problems of studying molecular pathways, but it is a more generalized model, which overcomes the deficiencies of current *conditional covariance selection* methods. Diagnostic prediction is another field of interest for cancer study. When a prediction model is built to find enriched biological groups for a case-control factor associated with a cancer disease, we want to have a more robust binary classification method for heterogeneous tumors, of which the expression pattern, even though for deconvolved data, is affected by how cancerous tissues contribute. In our dissertation, we develop an integrated pipeline of statistical modeling methods, from deconvolution, to correlation and classification, to aid immensely resolving the cellular heterogeneity problems in cancer studies.

5.2 Perspectives

This thesis develops a couple of statistical modeling approaches to the cellular heterogeneity problems. We aim to promote the development of statistical methods for incorporating tumor heterogeneity, and emphasize that extensions based on our work would be of great interest for future research. We suggest several important directions that can be extended in future work.

First, the identifiability of tumor deconvolution problem is not trivial. As discussed in Chapter 2, tumor deconvolution models, with whichever assumed underlying distribution for deconvolved signals, cannot guarantee the identifiability for estimation of component proportions. For existing deconvolution models, it is difficult to obtain a global optimization solution. Actually the best local optimums we are able to obtain do not only depend on the initialization, but also on the selection of genes for deconvolution. Different subset of genes selected for deconvolution can lead to different proportion estimates. Due to those properties, gene selection plays a crucial role in deconvolution, even more important than modeling and optimization design in some case. All the current literature on deconvolution lacks consideration and discussion of gene selection. Our work discusses about this issue and provides several empirical strategies, but a systematic study of gene selection for deconvolution with strictly theoretical discussion is expected for future research.

Second, it is desirable to include a non-linear parameterization along with its inference scheme in our edge regression model. The relationship between edge strength and exogenous covariates in our proposed edge regression model can be allowed to be linear or non-linear. We provide the sampling scheme for the linear relationship in this work. In the future work, spline-based semi-parametric representations can be used to adapt the *conditional precision function*. A generalized mixed model for

spline functions can help to develop a sampling scheme with similar logics for inference of linear relationship.

Third, it would be of great value to integrate tumor deconvolution with graphical models. On one side, as demonstrated in Chapter 3, tumor deconvolution techniques all require an individually independence assumption for gene expression. This is a very strong assumption, and it is actually contrary to the co-regulation and co-expression fact of genes. Some neighboring genes tend to be correlated on expression in tumor samples. Current strategy for deconvolution lacks the technique to include correlation analysis into inference. On the other side, current edge regression model utilizes the tumor purity, which is estimated from deconvolution, to infer subject-level and population-level graph, but the bias caused by the inference of deconvolution, would be introduced to the construction of networks. This secondary bias is hard to be estimated. Therefore, an integrated model to combine deconvolution of tumor expression pattern and construction of subject-level gene networks will contribute significantly to both of those two aspects, and even could potentially help to solve the gene selection problem.

All the three directions following my thesis work will prominently pave a road for the future work of statistical modeling and provide an aid to develop more precise and powerful tool for cellular heterogeneity problems. In sum, we resolve the cellular heterogeneity problems by applying statistical approaches including deconvolution, Gaussian graphical models and logistic regression to help the genomic analyses of tumor samples. Cancer biology is still making faster and faster progress to expose more and more interesting findings. With such advances, I hope my thesis would motivate more in-depth studies and provide new insights into related research topics in cancer studies.

Appendix A

Cell type-specific Deconvolution of Heterogeneous Tumor Samples using Expression Data

A.1 Proof of Local Optimality for ICM Algorithm

In these iterative steps of ICM, the complete likelihood is updated by searching conditional modes. It never decreases at any iteration and eventual convergence to the local maximum is guaranteed.

We organize the descriptive proof of local optimality as follows. In the t-th iteration, we search the maximum $\{\pi_1^{(t)}, \pi_2^{(t)}\}_i$ of $f(\{\pi_1, \pi_2\}_i; \{y_{ig}\}_{g=1}^G, \{\mu_T^{(t-1)}, \sigma_T^{(t-1)}\}_{g=1}^G)$ for each sample $i = 1, \cdots, S$ by using a golden search and successive parabolic interpolation.

Our complete likelihood function can be expressed as a product:

$$L(\{\pi_1, \pi_2\}_i^S, \{\mu_T, \sigma_T\}_{g=1}^G) = \prod_{i=1}^{S} f(\{\pi_1, \pi_2\}_i, \{y_{ig}\}_{g=1}^G, \{\mu_T, \sigma_T\}_{g=1}^G). \tag{A.1}$$

Then, this sample-wise maximum leads to

$$L(\{\pi_1^{(t)}, \pi_2^{(t)}\}_i^S, \{\mu_T^{(t-1)}, \sigma_T^{(t-1)}\}_{g=1}^G) \geq L(\{\pi_1^{(t-1)}, \pi_2^{(t-1)}\}_i^S, \{\mu_T^{(t-1)}, \sigma_T^{(t-1)}\}_{g=1}^G). \tag{A.2}$$

Similarly, we search the maximum $\{\mu_T^{(t)}, \sigma_T^{(t)}\}_g$ of $f(\{\mu_T, \sigma_T\}_g; \{y_{ig}\}_{i=1}^S, \{\pi_1^{(t)}, \pi_2^{(t)}\}_{i=1}^S)$ for each gene $i = 1, \cdots, G$. Our complete likelihood function can be alternatively expressed as a product:

$$L(\{\pi_1,\pi_2\}_i^S,\{\mu_T,\sigma_T\}_{g=1}^G)=\prod_{g=1}^{G}L(\{\mu_T,\sigma_T\}_g,\{y_{ig}\}_{i=1}^S,\{\pi_1,\pi_2\}_{i=1}^S). \tag{A.3}$$

Then, this genome-wise maximum leads to

$$L(\{\pi_1^{(t)},\pi_2^{(t)}\}_i^S,\{\mu_T^{(t)},\sigma_T^{(t)}\}_{g=1}^G)\geq L(\{\pi_1^{(t)},\pi_2^{(t)}\}_i^S,\{\mu_T^{(t-1)},\sigma_T^{(t-1)}\}_{g=1}^G). \tag{A.4}$$

Therefore we have a new inequality by combining the two aforementioned inequalities:

$$L(\{\pi_1^{(t)},\pi_2^{(t)}\}_i^S,\{\mu_T^{(t)},\sigma_T^{(t)}\}_{g=1}^G)\geq L(\{\pi_1^{(t-1)},\pi_2^{(t-1)}\}_i^S,\{\mu_T^{(t-1)},\sigma_T^{(t-1)}\}_{g=1}^G). \tag{A.5}$$

Thus, assuming we start with a guess for vector $\{\mu_{Tg}^{(0)},\sigma_{Tg}^{(0)}\}_g$ for each $g=1,\cdots,G$ for a local maximum of our complete likelihood L, and considering the sequence $\{\{\pi_1^{(0)},\pi_2^{(0)}\}_i^S,\{\mu_T^{(0)},\sigma_T^{(0)}\}_{g=1}^G\},\{\{\pi_1^{(1)},\pi_2^{(1)}\}_i^S,\{\mu_T^{(1)},\sigma_T^{(1)}\}_{g=1}^G\},\{\{\pi_1^{(2)},\pi_2^{(2)}\}_i^S,\{\mu_T^{(2)},\sigma_T^{(2)}\}_{g=1}^G\},\cdots$, we have

$$\begin{aligned}L(\{\pi_1^{(0)},\pi_2^{(0)}\}_i^S,\{\mu_T^{(0)},\sigma_T^{(0)}\}_{g=1}^G)&\leq L(\{\pi_1^{(1)},\pi_2^{(1)}\}_i^S,\{\mu_T^{(1)},\sigma_T^{(1)}\}_{g=1}^G)\\&\leq L(\{\pi_1^{(2)},\pi_2^{(2)}\}_i^S,\{\mu_T^{(2)},\sigma_T^{(2)}\}_{g=1}^G)\leq\cdots\end{aligned} \tag{A.6}$$

The sequence $\{\{\pi_1^{(t)},\pi_2^{(t)}\}_i^S,\{\mu_T^{(t)},\sigma_T^{(t)}\}_{g=1}^G\}$ should converge to the desired local optimum.

A.2 Supplemental Tables

Table A.1 : Summary of datasets GEO19830 with the mixture proportions (%) of rat liver, brain and lung, three of which are isolated as pure type tissue.

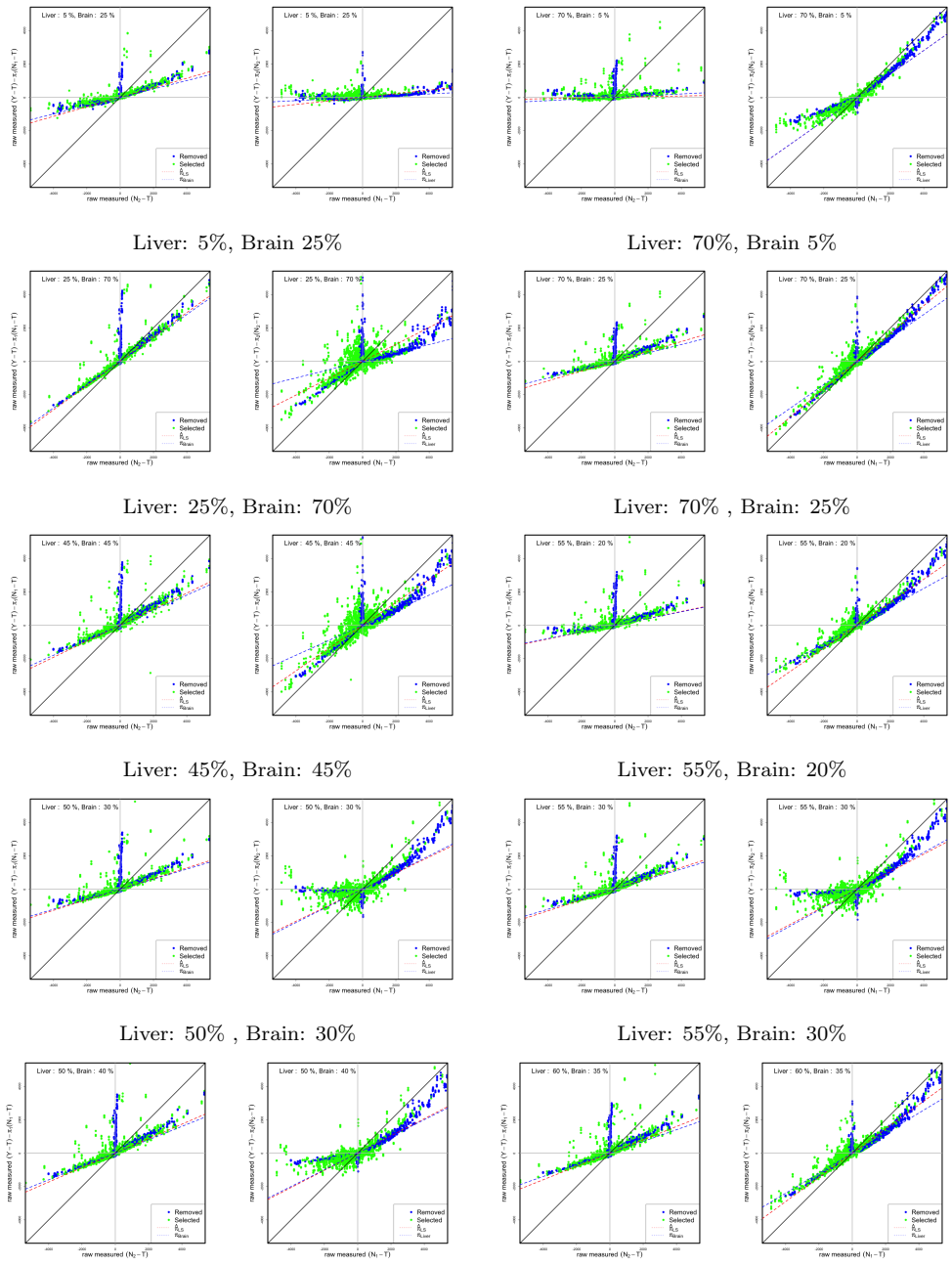
Mixture	Number of Technical Replicate	Tissue Type	Liver	Brain	Lung
1	3	Pure	100	0	0
2	3	Pure	0	100	0
3	3	Pure	0	0	100
4	3	Mixed	5	25	70
5	3	Mixed	70	5	25
6	3	Mixed	25	70	5
7	3	Mixed	70	25	5
8	3	Mixed	45	45	10
9	3	Mixed	55	20	25
10	3	Mixed	50	30	20
11	3	Mixed	55	30	15
12	3	Mixed	50	40	10
13	3	Mixed	60	35	5

Table A.2 : Summary of datasets in RNA-seq mixed cell lines experiment with the mixture proportions (%) of lung adenocarcinoma in humans (H1092), cancer-associated fibroblasts (CAFs) and tumor infiltrating lymphocytes (TIL), three of which are isolated as pure type tissue.

Mixture	Number of Technical Replicate	Tissue Type	H1092	CAF	TIL
1	2	Pure	100	0	0
2	2	Pure	100	0	0
3	2	Pure	100	0	0
4	2	Pure	0	100	0
5	2	Pure	0	100	0
6	2	Pure	0	100	0
7	2	Pure	0	0	100
8	2	Pure	0	0	100
9	2	Pure	0	0	100
10	2	Mixed	45.6	50.8	3.6
11	2	Mixed	45.6	50.8	3.6
12	2	Mixed	45.6	50.8	3.6
13	2	Mixed	61.9	35.6	2.5
14	2	Mixed	61.9	35.6	2.5
15	2	Mixed	61.9	35.6	2.5
16	2	Mixed	29.6	68	2.4
17	2	Mixed	29.6	68	2.4
18	2	Mixed	29.6	68	2.4
19	2	Mixed	43.2	49.7	7.1
20	2	Mixed	43.2	49.7	7.1

21	2	Mixed	43.2	49.7	7.1
22	2	Mixed	63	36.2	0.9
23	2	Mixed	63	36.2	0.9
24	2	Mixed	63	36.2	0.9
26	2	Mixed	30	69.1	0.8
27	2	Mixed	30	69.1	0.8
28	2	Mixed	30	69.1	0.8
29	2	Mixed	81.9	17.7	0.4
30	2	Mixed	81.9	17.7	0.4
31	2	Mixed	81.9	17.7	0.4
32	2	Mixed	93.6	6	0.4
33	2	Mixed	93.6	6	0.4

A.3 Supplemental Figures



Liver: 5%, Brain 25%

Liver: 70%, Brain 5%

Liver: 25%, Brain: 70%

Liver: 70% , Brain: 25%

Liver: 45%, Brain: 45%

Liver: 55%, Brain: 20%

Liver: 50% , Brain: 30%

Liver: 55%, Brain: 30%

Liver: 50%, Brain: 40%

Liver: 60%, Brain: 35%

Figure A.1 : Scatter plots of $Y_{ig} - T_{ig} - \pi_{2,i}(\bar{N}_{2,g} - T_{ig})$ versus $\bar{N}_{1,g} - T_{ig}$ and $Y_{ig} - T_{ig} - \pi_{1,i}(\bar{N}_{1,g} - T_{ig})$ versus $\bar{N}_{2,g} - T_{ig}$ for raw measured data at 10 different mixture ratios. Red dash line denotes the fitted regression coefficient for all probes by least squares; blue dash line denotes the truth purity; blue dots denote the probes we remove; green dots denote the remaining probes, from which the expression level of at least two of three tissues measure above 2^7 for deconvolution. If the linearity holds, the fitted line by regression on green dots should be approximate the line with slope equal to the true proportion.)

Appendix B

Bayesian Edge Regression for Undirected Graphical Model Accounting for Biological Heterogeneity

B.1 Summary of Notation

The context of edge regression enables the use of index set for exogenous covariates, which makes it complicated in notation. We summarize the notation of random vectors and their matrix form.

Table B.1 : Summary of notation

Symbol	Description
s	index of exogenous covariate
i, j	index of vertex
n	index of sample
q	number of exogenous covariates
p	number of vertices
N	sample size
β_s^{ij}	edge regression coefficient of s-th covariate for edge (i, j)
$\boldsymbol{\beta^{ij}} = (\beta_1^{ij}, \beta_2^{ij}, \cdots, \beta_q^{ij})^T$	q-dimensional random vector of β_s^{ij}
$\boldsymbol{\beta} = (\boldsymbol{\beta^{1,2}}, \boldsymbol{\beta^{1,3}}, \cdots, \boldsymbol{\beta^{p-1,p}})$	$q \times \frac{p(p-1)}{2}$-matrix of $\boldsymbol{\beta^{ij}}$
$X_{s,n}$	the s-th covariate of sample n
$\boldsymbol{X_n} = (X_{1,n}, X_{2,n}, \cdots, X_{q,n})^T$	q-dimensional random vector of exogenous covariates for sample n
$\boldsymbol{X} = (\boldsymbol{X_1^T}, \boldsymbol{X_2^T}, \cdots, \boldsymbol{X_N^T})^{\boldsymbol{T}}$	$N \times q$-matrix of exogenous covariates
Y_n^i	random variable of vertex i in graph for sample n
$\boldsymbol{Y_n} = (Y_n^1, Y_n^2, \cdots, Y_n^p)^T$	p-dimensional random vector for sample n
$\boldsymbol{Y} = (\boldsymbol{Y_1^T}, \boldsymbol{Y_2^T}, \cdots, \boldsymbol{Y_N^T})^{\boldsymbol{T}}$	$N \times p$-matrix of observed data
ψ_s^{ij}	scale parameter of normal prior for the s-th covariate of edge (i, j)
$\boldsymbol{\psi^{ij}} = diag(\psi_1^{ij}, \psi_2^{ij}, \cdots, \psi_q^{ij})$	$q \times q$-matrix of scale parameter of posterior probability for edge (i, j)
$\widetilde{\boldsymbol{\mu}}^{\boldsymbol{ij}}$, $\widetilde{\boldsymbol{\Sigma}}^{\boldsymbol{ij}}$	parameter of normal prior for $\boldsymbol{\beta^{ij}}$
$S_{1,n}$	element of calculation for $\widetilde{\boldsymbol{\mu}}^{\boldsymbol{ij}}$, $\widetilde{\boldsymbol{\Sigma}}^{\boldsymbol{ij}}$
$S_{1,n}$	element of calculation for $\widetilde{\boldsymbol{\mu}}^{\boldsymbol{ij}}$, $\widetilde{\boldsymbol{\Sigma}}^{\boldsymbol{ij}}$
$\boldsymbol{S_1} = diag(S_{1,1}, S_{1,2}, \cdots, S_{1,N})$	$N \times N$-matrix of $S_{1,n}$
$\boldsymbol{S_2} = \{S_{2,1}, S_{2,2}, \cdots, S_{2,N}\}^T$	N-dimensional vector of $S_{2,n}$

B.2 Details and Derivation of MCMC Sampling

In equation 3.3 we condition the precision matrix Ω on a set of exogenous covariates X. Instead of modeling $\gamma^{ij}(x)$, we model the conditional precision function on $\omega^{ij}(x)$ under the regression setting. Since the precision matrix is symmetric, we have $\omega^{ji}(.) = \omega^{ij}(.)$. Hence, we can coerce these two functions to have the same form in the sampling scheme. When we regress $\omega^{ij}(.)$ on X in a linear setting :

$$\omega^{ij}(x) = \sum_{s=1}^{q} \beta_s^{ij} X_s \tag{B.1}$$

We have $\beta_s^{ij} = \beta_s^{ji}$ for every $i \neq j$. Then we have a complete likelihood given by equation 3.4. In the context of using normal-gamma shrinkage prior, the posterior distribution of all parameters in our model can be updated through a Gibbs sampling scheme. We will also follow the normal-gamma prior paper (Griffin et al., 2010) to talk about updating hyper-parameters for normal-gamma prior through a Metropolis-Hasting step in our model.

Update β^{ij} for every pair $(i,j), i < j$ For $\boldsymbol{\beta^{ij}} = \{\beta_s^{ij}\}^S$ of any given pair of vertex (i,j), we derive the full conditional by

$$
\begin{aligned}
f(\boldsymbol{\beta^{ij}}|.) &\propto f(\boldsymbol{Y^i}|\boldsymbol{Y^{-i}}, \{\beta_s^{i,-(i,j)}\}_{s=1}^q, \{\beta_s^{i,j}\}_{s=1}^q, \omega^{i,i}, \{X_{s,n}\}_{s=1,n=1}^{q,N}) \\
&\times f(\boldsymbol{Y^j}|\boldsymbol{Y^{-j}}, \{\beta_s^{j,-(i,j)}\}_{s=1}^q, \{\beta_s^{i,j}\}_{s=1}^q, \omega^{j,j}, \{X_{s,n}\}_{s=1,n=1}^{q,N}) \times f(\boldsymbol{\beta^{i,j}}|\boldsymbol{\psi^{i,j}}) \\
&\propto \exp\{-\frac{1}{2}\sum_{n=1}^{N}[\frac{(Y_n^i + \frac{\sum_{k\neq i}^{p}\sum_{s=1}^{q}\beta_s^{ik}X_{s,n}Y_n^k}{\omega^{ii}})^2}{(\omega^{ii})^{-1}} + \frac{(Y_n^j + \frac{\sum_{k\neq j}^{p}\sum_{s=1}^{q}\beta_s^{jk}X_{s,n}Y_n^k}{\omega^{jj}})^2}{(\omega^{jj})^{-1}}]\} \\
&\times f(\boldsymbol{\beta^{i,j}}|\boldsymbol{\psi^{i,j}}) \\
&\propto \exp\{-\frac{1}{2}\sum_{n=1}^{N}[2Y_n^i\sum_{k\neq i}^{p}\sum_{s=1}^{q}\beta_s^{ik}X_{s,n}Y_n^k + \frac{(\sum_{k\neq i}^{p}\sum_{s=1}^{q}\beta_s^{ik}X_{s,n}Y_n^k)^2}{\omega^{ii}} \\
&+ 2Y_n^j\sum_{k\neq j}^{p}\sum_{s=1}^{q}\beta_s^{jk}X_{s,n}Y_n^k + \frac{(\sum_{k\neq j}^{p}\sum_{s=1}^{q}\beta_s^{jk}X_{s,n}Y_n^k)^2}{\omega^{jj}}]\} \times f(\boldsymbol{\beta^{i,j}}|\boldsymbol{\psi^{i,j}}) \\
&\propto \exp\{-\frac{1}{2}\sum_{n=1}^{N}[4Y_n^iY_n^j\sum_{s=1}^{q}\beta_s^{ij}X_{s,n} \\
&+ \frac{2\sum_{k\neq(i,j),s}\beta_s^{ik}X_{s,n}Y_n^k\sum_{s=1}^{q}\beta_s^{ij}X_{s,n}Y_n^j + (\sum_{s=1}^{q}\beta_s^{ij}X_{s,n}Y_n^j)^2}{\omega^{ii}} \\
&+ \frac{2\sum_{k\neq(i,j),s}\beta_s^{jk}X_{s,n}Y_n^k\sum_{s=1}^{q}\beta_s^{ij}X_{s,n}Y_n^i + (\sum_{s=1}^{q}\beta_s^{ij}X_{s,n}Y_n^i)^2}{\omega^{jj}}]\} \times f(\boldsymbol{\beta^{i,j}}|\boldsymbol{\psi^{i,j}}) \\
&\propto f(\boldsymbol{\beta^{i,j}}|\boldsymbol{\psi^{i,j}}) \times \exp\{-\frac{1}{2}\sum_{n=1}^{N}[(\sum_{s=1}^{q}\beta_s^{ij}X_{s,n})^2(\frac{(Y_n^j)^2}{\omega^{ii}} + \frac{(Y_n^i)^2}{\omega^{jj}}) \\
&+ 2\sum_{s=1}^{q}\beta_s^{ij}X_{s,n}(2Y_n^iY_n^j + \frac{\sum_{k\neq(i,j),s}\beta_s^{ik}X_{s,n}Y_n^kY_n^j}{\omega^{ii}} + \frac{\sum_{k\neq(i,j),s}\beta_s^{jk}X_{s,n}Y_n^kY_n^i}{\omega^{jj}})]\}
\end{aligned}
\tag{B.2}
$$

In the equation above, we simplify it by denoting:

$$
S_{1,n} = \frac{(Y_n^j)^2}{\omega^{ii}} + \frac{(Y_n^i)^2}{\omega^{jj}} \tag{B.3}
$$

and

$$
S_{2,n} = 2Y_n^iY_n^j + \sum_{k\neq(i,j),s}\beta_s^{ik}X_{s,n}Y_n^k\frac{Y_n^j}{\omega^{ii}} + \sum_{k\neq(i,j),s}\beta_s^{jk}X_{s,n}Y_n^k\frac{Y_n^i}{\omega^{jj}} \tag{B.4}
$$

For simplification of notation we have:

$$\sum_{k\neq(i,j),s} \beta_s^{ik} X_{s,n} Y_n^k = \boldsymbol{X_n^T \beta^{i,-j} Y_n^{-(i,j)}}$$
$$\sum_{k\neq(i,j),s} \beta_s^{jk} X_{s,n} Y_n^k = \boldsymbol{X_n^T \beta^{j,-i} Y_n^{-(j,i)}}. \tag{B.5}$$

, where $\boldsymbol{\beta^{i,-j}}$ corresponds to the columns that include i but not j in the superscript in $\boldsymbol{\beta}$, and $\boldsymbol{Y_n^{-(i,j)}}$ corresponds to the remaining elements in the vector $\boldsymbol{Y_n}$ after removing i-th and j-th element. This way, we rewrite:

$$S_{2,n} = 2Y_n^i Y_n^j + \boldsymbol{X_n}^T \boldsymbol{\beta^{i,-j} Y_n^{-(i,j)}} \frac{\boldsymbol{Y_n^j}}{\omega^{ii}} + \boldsymbol{X_n}^T \boldsymbol{\beta^{j,-i} Y_n^{-(j,i)}} \frac{\boldsymbol{Y_n^i}}{\omega^{jj}} \tag{B.6}$$

We also have:

$$\sum_{s=1}^{q} \beta_s^{ij} X_{s,n} = (\boldsymbol{\beta^{ij}})^T \boldsymbol{X_n}$$
$$(\sum_{s=1}^{q} \beta_s^{ij} X_{s,n})^2 = (\boldsymbol{\beta^{ij}})^T \boldsymbol{X_n X_n}^T \boldsymbol{\beta^{ij}} \tag{B.7}$$

Hence, we have:

$$\begin{aligned} B.2 &= f(\boldsymbol{\beta^{i,j}}|\boldsymbol{\psi^{i,j}}) \times \exp\{-\frac{1}{2}\sum_{n=1}^{N}[(\boldsymbol{\beta^{ij}})^T \boldsymbol{X_n X_n}^T \boldsymbol{\beta^{ij}} S_{1,n} + 2(\boldsymbol{\beta^{ij}})^T \boldsymbol{X_n} S_{2,n}]\} \\ &\propto \exp -\frac{1}{2}[(\boldsymbol{\beta^{ij}})^T (\boldsymbol{\psi^{ij}})^{-1} \boldsymbol{\beta^{ij}} + \sum_{n=1}^{N}((\boldsymbol{\beta^{ij}})^T \boldsymbol{X_n} S_{1,n} \boldsymbol{X_n}^T \boldsymbol{\beta^{ij}} + 2(\boldsymbol{\beta^{ij}})^T \boldsymbol{X_n} S_{2,n})] \\ &= \exp -\frac{1}{2}[(\boldsymbol{\beta^{ij}})^T (\boldsymbol{\psi^{ij}})^{-1} \boldsymbol{\beta^{ij}} + ((\boldsymbol{\beta^{ij}})^T \sum_{n=1}^{N} \boldsymbol{X_n} S_{1,n} \boldsymbol{X_n}^T \boldsymbol{\beta^{ij}} + 2(\boldsymbol{\beta^{ij}})^T \boldsymbol{X_n} S_{2,n})] \\ &= \exp -\frac{1}{2}[((\boldsymbol{\beta^{ij}})^T (\sum_{n=1}^{N} \boldsymbol{X_n} S_{1,n} \boldsymbol{X_n}^T + (\boldsymbol{\psi^{ij}})^{-1}) \boldsymbol{\beta^{ij}} + 2(\boldsymbol{\beta^{ij}})^T \sum_{n=1}^{N} \boldsymbol{X_n} S_{2,n})] \end{aligned} \tag{B.8}$$

According to equation B.8, we have $\boldsymbol{\beta^{ij}}|. \sim N(\boldsymbol{\widetilde{\mu}^{ij}}, \boldsymbol{\widetilde{\Sigma}^{ij}})$,where

$$\begin{aligned}\widetilde{\boldsymbol{\mu}}^{\boldsymbol{ij}} &= -(\sum_{n=1}^{N} \boldsymbol{X_n} S_{1,n} \boldsymbol{X_n}^T + (\boldsymbol{\psi}^{\boldsymbol{ij}})^{-1})^{-1} \sum_{n=1}^{N} \boldsymbol{X_n} S_{2,n} \\ \widetilde{\boldsymbol{\Sigma}}^{\boldsymbol{ij}} &= (\sum_{n=1}^{N} \boldsymbol{X_n} S_{1,n} \boldsymbol{X_n}^T + (\boldsymbol{\psi}^{\boldsymbol{ij}})^{-1})^{-1}\end{aligned} \tag{B.9}$$

For further simplification, we formulate S_1 and S_2 in a matrix form according to Section B.1. The calculation of element-wise $\boldsymbol{X_n}^T \boldsymbol{\beta}^{\boldsymbol{i},-\boldsymbol{j}} \boldsymbol{Y_n}^{-(\boldsymbol{i},\boldsymbol{j})}$ is complicated in $\boldsymbol{S_2}$. We calculate this component vector through matrix algebra, where

$$\{\boldsymbol{X_n}^T \boldsymbol{\beta}^{\boldsymbol{i},-\boldsymbol{j}} \boldsymbol{Y_n}^{-(\boldsymbol{i},\boldsymbol{j})}\}_{n=1}^{N} = diag(\boldsymbol{X}\boldsymbol{\beta}^{\boldsymbol{i},-\boldsymbol{j}} (\boldsymbol{Y}^{-(\boldsymbol{i},\boldsymbol{j})})^T) \tag{B.10}$$

Then we can express (11) as:

$$\begin{aligned}\widetilde{\boldsymbol{\mu}}^{\boldsymbol{ij}} &= -(\boldsymbol{X}^T \boldsymbol{S_1} \boldsymbol{X} + (\boldsymbol{\psi}^{\boldsymbol{ij}})^{-1})^{-1} \boldsymbol{X}^T \boldsymbol{S_2} \\ \widetilde{\boldsymbol{\Sigma}}^{\boldsymbol{ij}} &= (\boldsymbol{X}^T \boldsymbol{S_1} \boldsymbol{X} + (\boldsymbol{\psi}^{\boldsymbol{ij}})^{-1})^{-1}\end{aligned} \tag{B.11}$$

Update $\omega^{ii}, i = 1, 2, \cdots, p$ By setting $f(\omega^{ii}) \propto 1$, we derive the full conditional by:

$$\begin{aligned} f(\omega^{ii}|.) &\propto f(\boldsymbol{Y^i}|\boldsymbol{X}, \boldsymbol{Y^{-i}}, \{\beta_s^{i,-(i,j)}\}_{s=1}^{q}, \{\beta_s^{i,j}\}_{s=1}^{q}, \omega^{i,i}, X_{s,n}) \times f(\omega^{i,i}) \\ &\propto (\omega^{ii})^{\frac{n}{2}} \exp\{-\frac{1}{2} \sum_{n=1}^{N} [\frac{(Y_n^i + \frac{\sum_{k\neq i}^{p} \sum_{s=1}^{q} \beta_s^{ik} X_{s,n} Y_n^k}{\omega^{ii}})^2}{(\omega^{ii})^{-1}}]\} \\ &\propto (\omega^{ii})^{\frac{n}{2}} \exp\{-\frac{1}{2} \sum_{n=1}^{N} [(Y_n^i)^2 \omega^{ii} + \frac{(\sum_{k\neq i}^{p} \sum_{s=1}^{q} \beta_s^{ik} X_{s,n} Y_n^k)^2}{\omega^{ii}}]\} \\ &\propto (\omega^{ii})^{\frac{n}{2}} \exp\{-\frac{1}{2} [\omega^{ii} \sum_{n=1}^{N} (Y_n^i)^2 + \sum_{n=1}^{N} \frac{\boldsymbol{X_n}^T \boldsymbol{\beta^{i\cdot}} \boldsymbol{Y_n^{-i}} (\boldsymbol{Y_n^{-i}})^T (\boldsymbol{\beta^{i\cdot}})^T \boldsymbol{X_n}}{\omega^{ii}}]\} \\ &\propto (\omega^{ii})^{\frac{n}{2}} \exp\{-\frac{1}{2} [\omega^{ii} \sum_{n=1}^{N} (Y_n^i)^2 + \frac{diag(\boldsymbol{X}\boldsymbol{\beta^{i\cdot}} (\boldsymbol{Y^{-i}})^T) diag(\boldsymbol{X}\boldsymbol{\beta^{i\cdot}} (\boldsymbol{Y^{-i}})^T)^T}{\omega^{ii}}]\} \end{aligned} \tag{B.12}$$

Hence, $\omega^{ii}|. \sim GIG(\frac{n}{2}+1, \sum_{n=1}^{N} (Y_n^i)^2, diag(\boldsymbol{X}\boldsymbol{\beta^{i\cdot}} (\boldsymbol{Y^{-i}})^T) diag(\boldsymbol{X}\boldsymbol{\beta^{i\cdot}} (\boldsymbol{Y^{-i}})^T)^T)$, where GIG is the Generalized Inverse Gaussian distribution.

Update ψ_s^{ij} We sample ψ_s^{ij} according to

$$f(\psi_s^{ij}|.) \propto f(\beta_s^{ij}|\psi_s^{ij}) \times f(\psi^{ij}|\lambda_s, \gamma) \tag{B.13}$$

That is equally, $\psi_s^{ij} \sim GIG(\lambda_s - \frac{1}{2}, 1/\gamma^2, (\beta_s^{ij})^2)$

Update λ_s and γ Instead of using cross validation technique to select λ and γ in normal-gamma prior, we choose to sample these parameters by specifying hyper-priors. We update λ_s and γ through a Metropolis-Hasting sampling method.

If we use $\pi(\lambda_s)$ to denote the prior of λ_s, we can have the full conditional of λ_s as

$$f(\lambda_s|\cdot) \propto \pi(\lambda_s) \frac{1}{(2\gamma^2)^{\frac{p(p-1)}{2}\lambda_s}(\Gamma(\lambda_s))^{\frac{p(p-1)}{2}}} (\prod_{i\neq j} \psi_s^{ij})^{\lambda_s} \tag{B.14}$$

, where we set $\pi(\lambda_s) \sim exp(1)$.

We have multiplicative random walk updates on λ_s through $\lambda_s^* = exp(\sigma_{\lambda_s}^2 z)\lambda_s$, where z satisfies a standard normal. $\sigma_{\lambda_s}^2$ is a tuning parameter for random walk and it is chosen so that the acceptance rate is around 20% to 30%. Then the acceptance function is given by:

$$min\left\{1, \frac{\lambda_s^*}{\lambda_s} \frac{\pi(\lambda_s^*)}{\pi(\lambda_s)} \frac{(2\gamma^2)^{\frac{p(p-1)}{2}\lambda_s^*}(\Gamma(\lambda_s^*))^{\frac{p(p-1)}{2}}}{(2\gamma^2)^{\frac{p(p-1)}{2}\lambda_s}(\Gamma(\lambda_s))^{\frac{p(p-1)}{2}}} (\prod_{i\neq j} \psi_s^{ij})^{\lambda_s^* - \lambda_s}\right\} \tag{B.15}$$

For the scale parameter γ, we follow the suggested setting in normal-gamma prior paper (Griffin et al., 2010), with $\sum_s \lambda_s \gamma^2 \sim Ga(2, \sum M_s))$. M_s is a hyper-parameter to approximately control the scale of $\lambda_s\gamma^2$ for the s-th covariate, so we provide a heuristic solution to it by calculating the mean square error between zero and off-diagonal elements in maximum likelihood estimator (MLE) of $\boldsymbol{\Sigma}$ corresponding to each s. $M_s = \frac{\sum_{1\leq i\leq j\leq p}(\hat{\Sigma}_{s,ij}^{-1})^2}{p(p-1)/2}$, where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{s}}^{-\mathbf{1}}$ is the inverse of the estimated Gaussian covariance matrix through MLE for samples considering the effect represented by the s-th covariate. When $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{s}}$ is singular, we can use

the estimated precision matrix $\hat{\boldsymbol{\Omega}}_{\boldsymbol{s}}$ through a regularization method, e.g. graphical lasso (Friedman et al., 2008), instead of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{s}}^{-\mathbf{1}}$. The derivation of M_s is discussed case by case in the section of simulation and case study.

Hence we have, $\gamma^{-2} \sim Ga(2 + qp(p-1)\lambda_s/2, \sum_s M_s/(2\sum_s \lambda_s) + \frac{1}{2}\sum_s \sum_{i \neq j} \psi_s^{ij})$.

B.3 Supplementary Tables

Table B.2 : Table of edge connectedness for the unions of the GH, Angiogenesis and Metabolic pathway in tumor graph, normal graph and common edges under $\kappa = 0.1$ (sorted by number of shared edges).

Symbol	Tumor	Normal	Common
IGFBP6	10	7	5
IL-1 beta	10	6	5
IGFBP-3	13	5	4
MIP-1 alpha	5	4	4
EGF	6	3	3
IGFBP4	11	3	3
IGFBP5	10	4	3
IL-6	6	4	3
ITAC	11	4	3
MIP-1 beta	7	6	3
SCF	7	6	3
VEGFR-2	12	5	3
ANG-1	7	2	2
Endoglin	9	5	2
FABP, adipocyte	7	3	2
Hepsin	16	4	2

IGFBP-2	7	6	2
IL-10	3	2	2
MMP-9, total	5	2	2
PDGF-BB	5	2	2
PECAM-1	9	3	2
PPP	3	2	2
Resistin	8	4	2
SDF-1	4	4	2
TNF-alpha	4	4	2
uPAR	10	4	2
VEGF	7	4	2
ACE	3	2	1
CA-15-3	6	1	1
CEACAM1	2	4	1
Decorin	10	2	1
FSH	2	5	1
G-CSF	1	2	1
GLP-1 total	10	3	1
HER-2	8	3	1
IFN-gamma	5	3	1
IL-1 alpha	5	4	1
IL-2	5	1	1
IL-8	8	3	1
Insulin	1	4	1
Kallikrein 5	2	1	1
KLK-7	2	1	1

Leptin	4	3	1
LH	3	1	1
MCP-1	6	2	1
MIF	5	1	1
MMP-3	4	2	1
MSLN	8	3	1
PLGF	8	3	1
TN-C	8	1	1
TNF-beta	4	4	1

Table B.3 : Table of edge connectedness for the Inflammation pathway in tumor graph, normal graph and common edges under $\kappa = 0.1$ (sorted by number of shared edges).

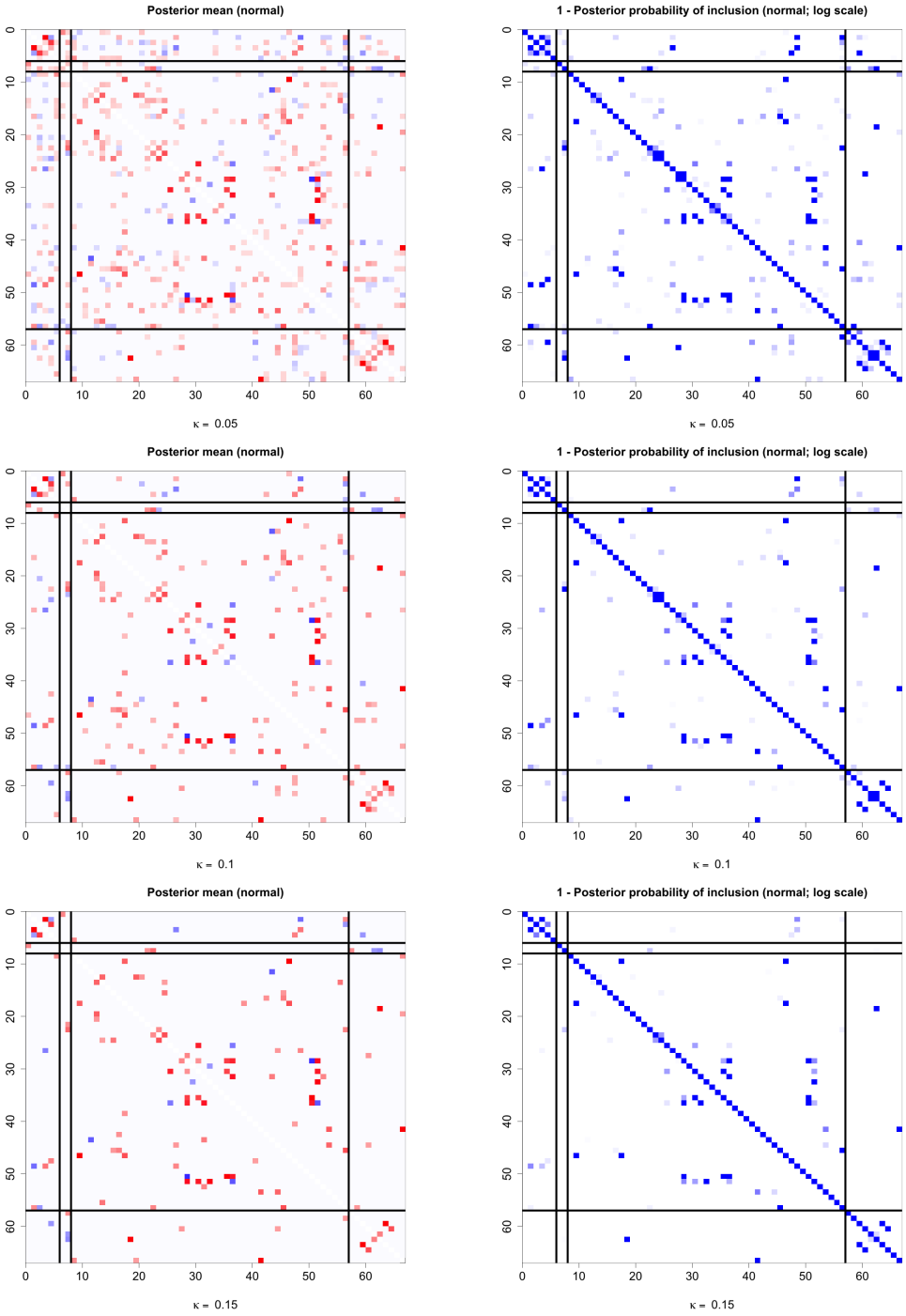
Symbol	Tumor	Normal	Common
MIP-1 beta	11	11	6
VCAM-1	11	7	6
MMP-3	9	6	5
BDNF	9	7	4
Factor VII	5	5	4
GM-CSF	9	9	4
IL-1 beta	10	7	4
IL-12p40	9	5	4
IL-3	7	8	4
MIP-1 alpha	8	5	4
TNF-beta	6	11	4
AAT	7	6	3
B2M	7	4	3

Fibrinogen	5	5	3
IL-1 alpha	13	3	3
IL-10	12	3	3
IL-1ra	10	8	3
IL-23	3	5	3
IL-6	5	8	3
TNF-alpha	6	5	3
IFN-gamma	7	3	2
IL-2	6	4	2
IL-7	8	9	2
IL-8	7	4	2
MMP-9	6	3	2
TIMP-1	10	4	2
TNFR2	3	3	2
VEGF	11	3	2
vWF	7	5	2
A2Macro	3	3	1
CRP	8	1	1
Eotaxin-1	4	4	1
Haptoglobin	3	4	1
IL-15	2	2	1
RANTES	9	2	1
SCF	7	3	1
VDBP	4	3	1

Table B.4 : Table of edge connectedness for the Immune pathway in tumor graph, normal graph and common edges under $\kappa = 0.1$ (sorted by number of shared edges).

Symbol	Tumor	Normal	Common
Thrombospondin-1	5	5	3
IL-16	4	2	2
IL-22	6	2	2
IP-10	6	3	2
MCP-2	7	5	2
MMP-9, total	5	4	2
MPIF-1	7	2	2
MPO	6	3	2
ANG-2	11	2	1
APRIL	9	2	1
AXL	6	3	1
CD40-L	2	3	1
DKK-1	2	2	1
Eotaxin-3	4	1	1
IL-6r	9	1	1
ITAC	8	2	1
MCP-4	5	2	1
MIF	4	3	1
MMP-1	3	2	1
TARC	5	1	1
TRAIL-R3	6	2	1

B.4 Supplementary Figures



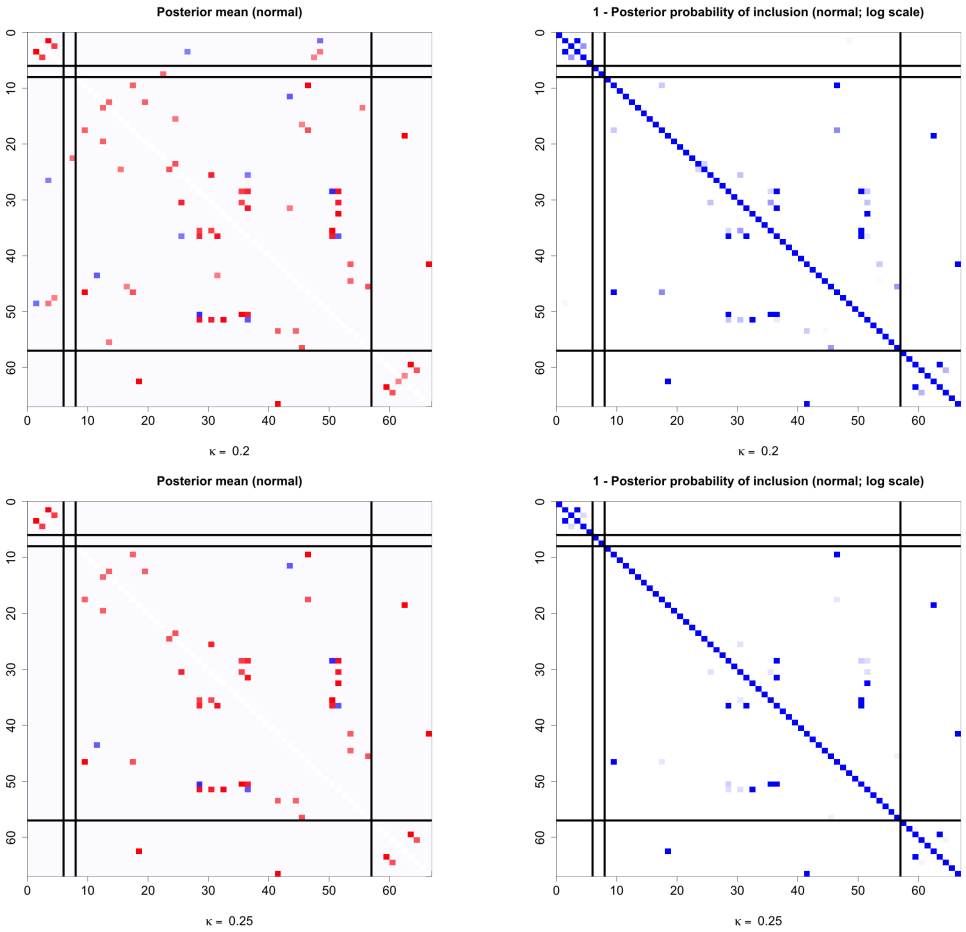
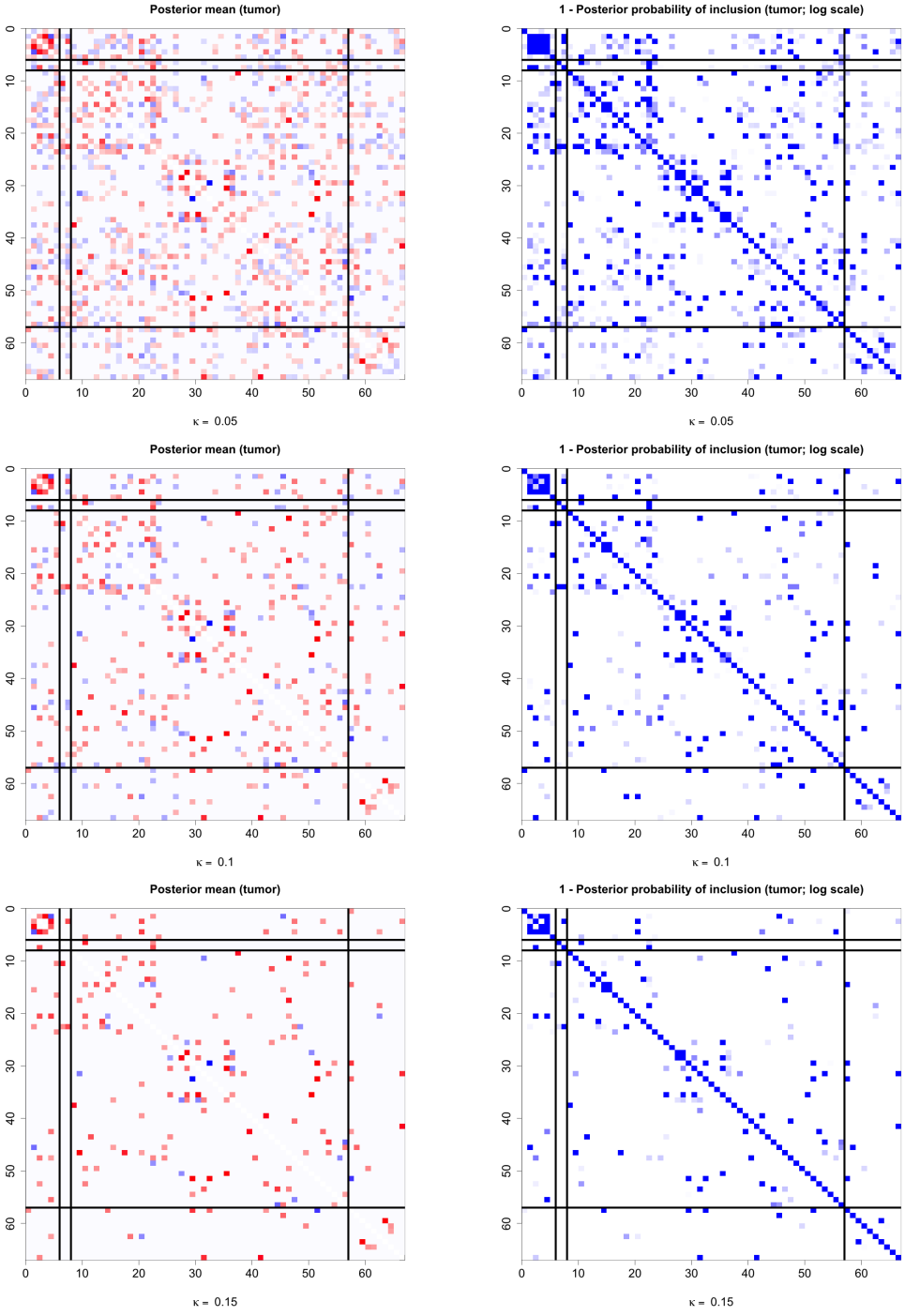


Figure B.1 : Heatmaps represent edge strength and 1 - PPI under different κ for normal graph. The white color indicates no connected edge. Left panel is for posterior mean calculated for partial correlation; right panel is for 1- Posterior Probability of Edge inclusion. On the left panel, blue indicates negative point estimates and red indicates positive point estimates (deeper color for larger absolute value). On the right panel, the degree of blue hue indicates the scale of local false discovery rate (deeper blue for smaller value) for each edge.


Posterior mean (tumor)
1 - Posterior probability of inclusion (tumor; log scale)
κ = 0.05
κ = 0.05
Posterior mean (tumor)
1 - Posterior probability of inclusion (tumor; log scale)
κ = 0.1
κ = 0.1
Posterior mean (tumor)
1 - Posterior probability of inclusion (tumor; log scale)
κ = 0.15
κ = 0.15

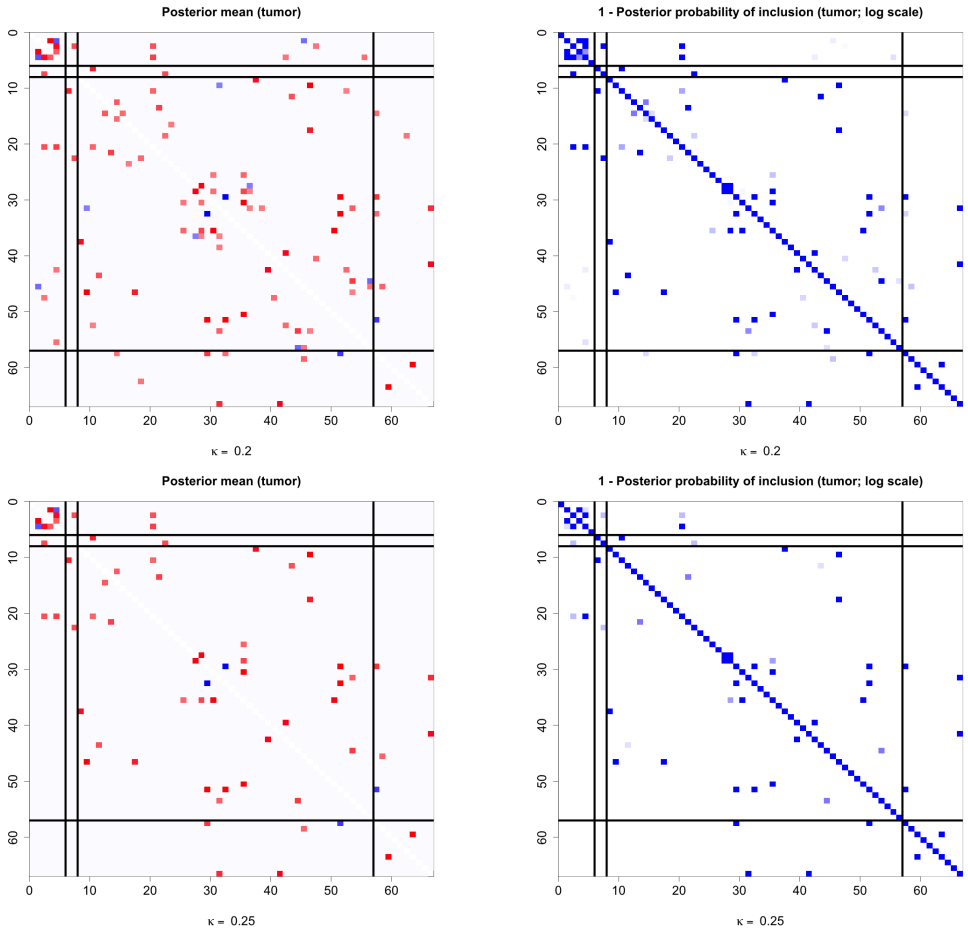
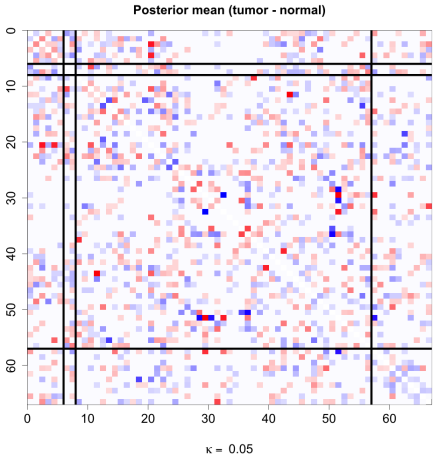
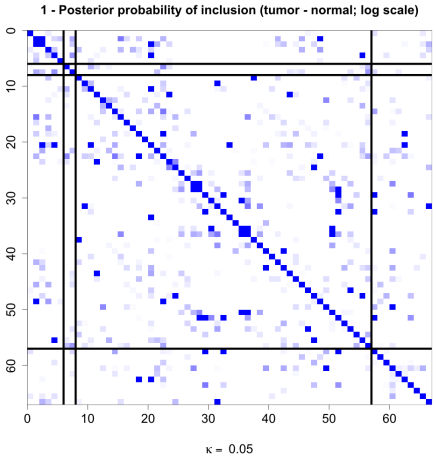
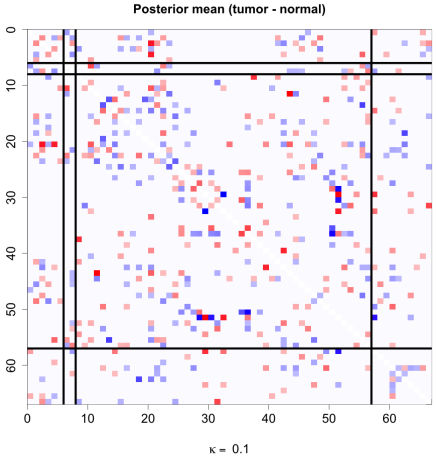
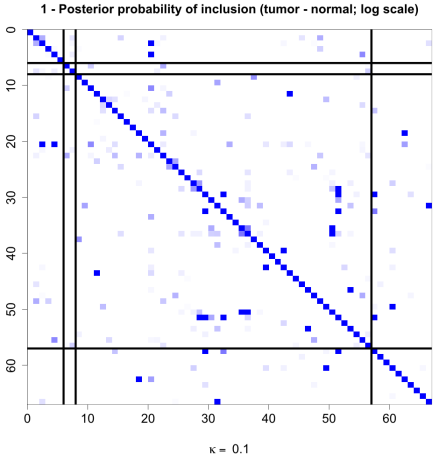


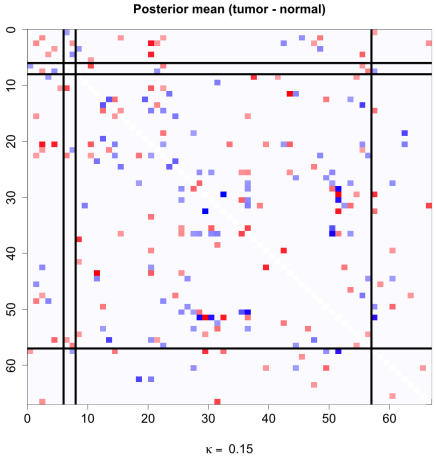
Figure B.2 : Heatmaps represent edge strength and 1 - PPI under different κ for tumor graph. The white color indicates no connected edge. Left panel is for posterior mean calculated for partial correlation; right panel is for 1- Posterior Probability of Edge inclusion. On the left panel, blue indicates negative point estimates and red indicates positive point estimates (deeper color for larger absolute value). On the right panel, the degree of blue hue indicates the scale of local false discovery rate (deeper blue for smaller value) for each edge.

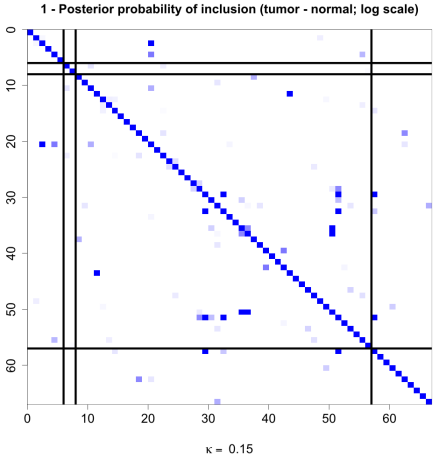

Posterior mean (tumor - normal)
κ = 0.05


1 - Posterior probability of inclusion (tumor - normal; log scale)
κ = 0.05


Posterior mean (tumor - normal)
κ = 0.1


1 - Posterior probability of inclusion (tumor - normal; log scale)
κ = 0.1


Posterior mean (tumor - normal)
κ = 0.15


1 - Posterior probability of inclusion (tumor - normal; log scale)
κ = 0.15

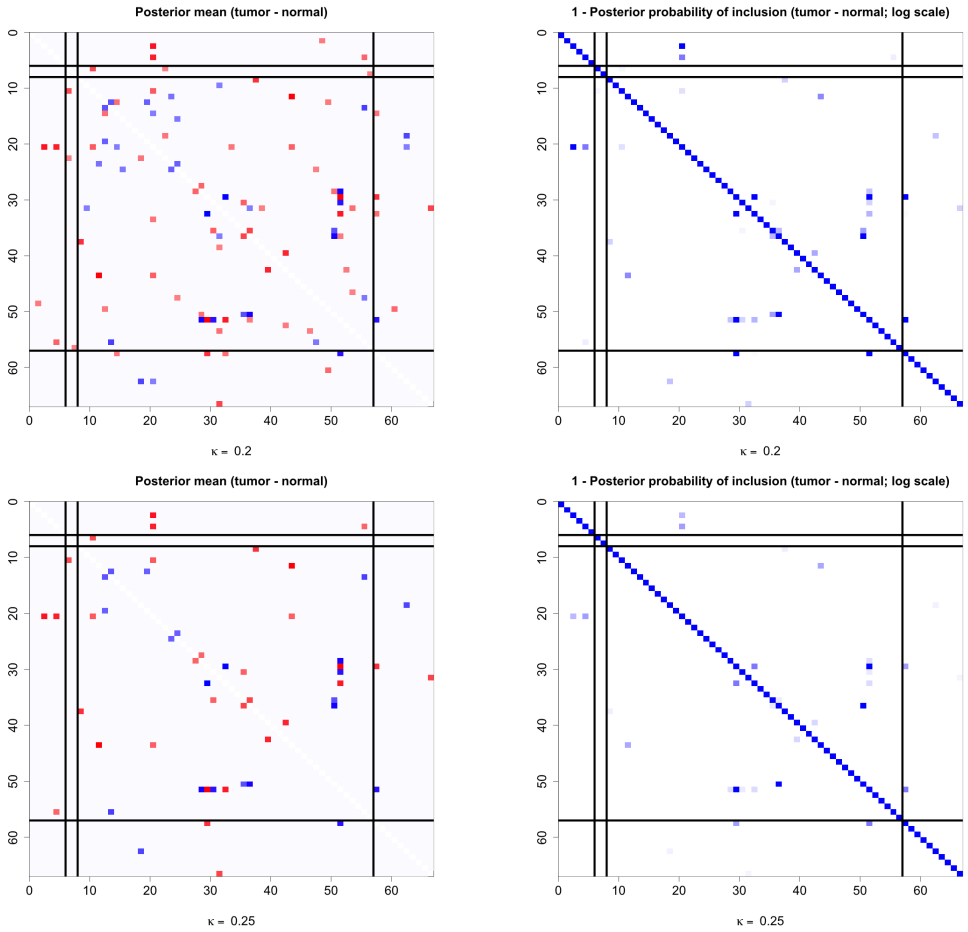


Figure B.3 : Heatmaps represent edge strength and 1 - PPI under different κ for differential edges between tumor and normal graph. The white color indicates no connected edge. Left panel is for posterior mean calculated for partial correlation; right panel is for 1- Posterior Probability of Edge inclusion. On the left panel, blue indicates negative point estimates and red indicates positive point estimates (deeper color for larger absolute value). On the right panel, the degree of blue hue indicates the scale of local false discovery rate (deeper blue for smaller value) for each edge.

Bibliography

Ahn, J., Yuan, Y., Parmigiani, G., Suraokar, M. B., Diao, L., Wistuba, I. I., and Wang, W. (2013). Demix: Deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, 29(15):1865–1871.

Aran, D., Sirota, M., and Butte, A. J. (2015). Systematic pan-cancer analysis of tumour purity. *Nature communications*, 6.

Aravalli, R. N., Steer, C. J., and Cressman, E. N. (2008). Molecular mechanisms of hepatocellular carcinoma. *Hepatology*, 48(6):2047–2063.

Bachtiary, B., Boutros, P. C., Pintilie, M., Shi, W., Bastianutto, C., Li, J.-H., Schwock, J., Zhang, W., Penn, L. Z., Jurisica, I., et al. (2006). Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity. *Clinical cancer research*, 12(19):5632–5640.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 259–302.

Bolstad, B., Irizarry, R., Åstrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Brent, R. P. (1973). *Algorithms for minimization without derivatives*. Courier Corporation.

Carroll, R. J. and Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 693–706.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., et al. (2012). Absolute quantification of somatic dna alterations in human cancer. *Nature biotechnology*, 30(5):413–421.

Carvalho, B., Bengtsson, H., Speed, T. P., and Irizarry, R. A. (2007). Exploration, normalization, and genotype calls of high-density oligonucleotide snp array data. *Biostatistics*, 8(2):485–499.

Cascone, T., Herynk, M. H., Xu, L., Du, Z., Kadara, H., Nilsson, M. B., Oborn, C. J., Park, Y.-Y., Erez, B., Jacoby, J. J., et al. (2011). Upregulated stromal egfr and vascular remodeling in mouse xenograft models of angiogenesis inhibitor–resistant human lung adenocarcinoma. *The Journal of clinical investigation*, 121(4):1313–1328.

Cawley, G. C. and Talbot, N. L. (2006). Gene selection in cancer classification using sparse logistic regression with bayesian regularization. *Bioinformatics*, 22(19):2348–2355.

Chen, Y., Sawyers, C. L., and Scher, H. I. (2008). Targeting the androgen receptor pathway in prostate cancer. *Current opinion in pharmacology*, 8(4):440–448.

Cheng, J., Levina, E., Wang, P., and Zhu, J. (2014). A sparse ising model with covariates. *Biometrics*, 70(4):943–953.

Chu, W., Ghahramani, Z., Falciani, F., and Wild, D. L. (2005). Biomarker discovery in microarray gene expression data with gaussian processes. *Bioinformatics*, 21(16):3385–3393.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397.

Dave, S. S., Wright, G., Tan, B., Rosenwald, A., Gascoyne, R. D., Chan, W. C., Fisher, R. I., Braziel, R. M., Rimsza, L. M., Grogan, T. M., Miller, T. P., LeBlanc, M., Greiner, T. C.,

Weisenburger, D. D., Lynch, J. C., Vose, J., Armitage, J. O., Smeland, E. B., Kvaloy, S., Holte, H., Delabie, J., Connors, J. M., Lansdorp, P. M., Ouyang, Q., Lister, T. A., Davies, A. J., Norton, A. J., Muller-Hermelink, H. K., Ott, G., Campo, E., Montserrat, E., Wilson, W. H., Jaffe, E. S., Simon, R., Yang, L., Powell, J., Zhao, H., Goldschmidt, N., Chiorazzi, M., and Staudt, L. M. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *New England Journal of Medicine*, 351(21):2159–2169. PMID: 15548776.

de Ridder, D., van der Linden, C. E., Schonewille, T., Dik, W. A., Reinders, M. J. T., van Dongen, J. J. M., and Staal, F. J. T. (2005). Purity for clarity: the need for purification of tumor cells in dna microarray studies. *Leukemia*, 19(4):618–627.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, pages 157–175.

Dettling, M. and Bühlmann, P. (2004). Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, 90(1):106–131.

Dhanasekaran, R., Bandoh, S., and Roberts, L. R. (2016). Molecular pathogenesis of hepatocellular carcinoma and impact of therapeutic advances. *F1000Research*, 5.

Dobson, A. J. and Barnett, A. (2008). *An introduction to generalized linear models*. CRC press.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87.

Emmert-Buck, M. R., Bonner, R. F., Smith, P. D., Chuaqui, R. F., et al. (1996). Laser capture microdissection. *Science*, 274(5289):998.

Farley, P. (2015). 'purity' of tumor samples may significantly bias genomic analyses. *UCSF News Center*.

Fattovich, G., Stroffolini, T., Zagni, I., and Donato, F. (2004). Hepatocellular carcinoma in cirrhosis: incidence and risk factors. *Gastroenterology*, 127(5):S35–S50.

Fridman, W. H., Pagès, F., Sautès-Fridman, C., and Galon, J. (2012). The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer*, 12(4):298–306.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., Wind, P., Zinzindohoué, F., Bruneval, P., Cugnenc, P.-H., Trajanoski, Z., Fridman, W.-H., and Pagès, F. (2006). Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*, 313(5795):1960–1964.

Gay, L., Baker, A.-M., and Graham, T. A. (2016). Tumour cell heterogeneity. *F1000Research*, 5.

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., Nair, V. S., Xu, Y., Khuong, A., Hoang, C. D., Diehn, M., B., W. R., K., P. S., and Alizadeh, A. A. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature medicine*, 21(8):938–945.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537.

Gong, T. and Szustakowski, J. D. (2013). Deconrnaseq: A statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data. *Bioinformatics*, 29(8):1083–1085.

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, 55(292):708–713.

Griffin, J. E., Brown, P. J., et al. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796.

Heppner, G. H. and Miller, B. E. (1983). Tumor heterogeneity: biological implications and therapeutic consequences. *Cancer and Metastasis Reviews*, 2(1):5–23.

Hoff, P. D. and Niu, X. (2012). A covariance regression model. *Statistica Sinica*, pages 729–753.

Hoti, F. and Sillanpää, M. (2006). Bayesian mapping of genotype× expression interactions in quantitative and qualitative traits. *Heredity*, 97(1):4–18.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, pages 46–60.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. biostatistics. *Biostatistics*, 4(2):249–264.

Junttila, M. R. and de Sauvage, F. J. (2013). Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature*, 501(7467):346–354.

Kalluri, R. and Zeisberg, M. (2006). Fibroblasts in cancer. *Nat Rev Cancer*, 6(5):392–401.

Kolar, M., Parikh, A. P., and Xing, E. P. (2010). On sparse nonparametric conditional covariance selection. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 559–566.

Kolar, M. and Xing, E. P. (2009). Sparsistent estimation of time-varying discrete markov random fields. *arXiv preprint arXiv:0907.2337*.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):R29.

Lawrence, I. and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268.

Leday, G. G., de Gunst, M., Kpogbezan, G. B., Van der Vaart, A. W., Van Wieringen, W. N., and Van de Wiel, M. A. (2015). Gene network reconstruction using global-local shrinkage priors. *arXiv preprint arXiv:1510.03771*.

Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J. C., Rodig, S., et al. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biology*, 17(1):174.

Li, Y., Campbell, C., and Tipping, M. (2002). Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, 18(10):1332–1339.

Liao, J. and Chin, K.-V. (2007). Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, 23(15):1945–1951.

Liebner, D. A., Huang, K., and Parvin, J. D. (2014). Mmad: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics*, 30(5):682–689.

Liotta, L. and Petricoin, E. (2000). Molecular profiling of human cancer. *Nature Reviews Genetics*, 1(1):48–56.

Liu, H., Chen, X., Wasserman, L., and Lafferty, J. D. (2010). Graph-valued regression. In *Advances in Neural Information Processing Systems*, pages 1423–1431.

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica sinica*, 12(1):31–46.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312.

Marjanovic, N. D., Weinberg, R. A., and Chaffer, C. L. (2013). Cell plasticity and heterogeneity in cancer. *Clinical chemistry*, 59(1):168–179.

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334.

McGuffey, E. J. and et al (2017). Bivariate auc.

Meacham, C. E. and Morrison, S. J. (2013). Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462.

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A., and Coombes, K. R. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, 64(2):479–489.

Network, C. G. A. et al. (2015a). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536):576–582.

Network, C. G. A. R. et al. (2015b). The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025.

Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50.

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2017). Bayesian graphical regression.

Nikooienejad, A., Wang, W., and Johnson, V. E. (2016). Bayesian variable selection for binary outcomes in high dimensional genomic studies using non-local priors. *Bioinformatics*, page btv764.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117.

Pages, F., Galon, J., Dieu-Nosjean, M.-C., Tartour, E., Sautes-Fridman, C., and Fridman, W.-H. (2009). Immune infiltration in human tumors: a prognostic factor that should not be ignored. *Oncogene*, 29(8):1093–1102.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2012). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*.

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174.

Quon, G., Haider, S., Deshwar, A. G., Cui, A., Boutros, P. C., and Morris, Q. (2013). Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med*, 5(3):29.

Quon, G. and Morris, Q. (2009). Isolate: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics*, 25(21):2882–2889.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rockafellar, R. T. (2015). *Convex analysis*. Princeton university press.

Rodríguez, G. (2007). Lecture notes on generalized linear models.

Sanyal, A. J., Yoon, S. K., and Lencioni, R. (2010). The etiology of hepatocellular carcinoma and consequences for treatment. *The oncologist*, 15(Supplement 4):14–22.

Sartor, M. A., Leikauf, G. D., and Medvedovic, M. (2009). Lrpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2):211–217.

Shen-Orr, S. S. and Gaujoux, R. (2013). Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current opinion in immunology*, 25(5):571–578.

Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M., Hastie, T., Sarwal, M. M., Davis, M. M., and Butte, A. J. (2010). Cell type-specific gene expression differences in complex tissues. *Nat Meth*, 7(4):287–289.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., De Longueville, F., Kawasaki, E. S., Lee, K. Y., et al. (2006). The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9):1151–1161.

Šimecková, M. (2005). Maximum weighted likelihood estimator in logistic regression.

Simon, R. (2003). Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British journal of cancer*, 89(9):1599–1604.

Su, X., Zhang, L., Zhang, J., Meric-Bernstam, F., and Weinstein, J. N. (2012). Purityest: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics*, 28(17):2265–2266.

Sun, X.-x. and Yu, Q. (2015). Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacologica Sinica*, 36(10):1219–1227.

Vandev, D. and Neykov*, N. (1998). About regression estimators with high breakdown point. *Statistics: A Journal of Theoretical and Applied Statistics*, 32(2):111–129.

Venables, W. N. and Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., Meehan, J., Li, X., Yang, L., Li, H., Labaj, P. P., Kreil, D. P., Megherbi, D., Gaj, S., Caiment, F., van Delft, J., Kleinjans, J., Scherer, A., Devanarayan, V., Wang, J., Yang, Y., Qian, H.-R., Lancashire, L. J., Bessarabova, M., Nikolsky, Y., Furlanello, C., Chierici, M., Albanese, D., Jurman, G., Riccadonna, S., Filosi, M., Visintainer, R., Zhang, K. K., Li, J., Hsieh, J.-H., Svoboda, D. L., Fuscoe, J. C., Deng, Y., Shi, L., Paules, R. S., Auerbach, S. S., and Tong, W. (2014). The

concordance between rna-seq and microarray data depends on chemical treatment and transcript abundance. *Nat Biotech*, 32(9):926–932.

Wang, N., Gong, T., Clarke, R., Chen, L., Shih, I.-M., Zhen Zhang, D. A. L., Xuan, J., and Wang, Y. (2015). Undo: a bioconductor r package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics*, 31(1):137–139.

Wang, X., Zidek, J. V., et al. (2005). Selecting likelihood weights by cross-validation. *The Annals of Statistics*, 33(2):463–500.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98(20):11462–11467.

West, N., Dattani, M., McShane, P., Hutchins, G., Grabsch, J., Mueller, W., Treanor, D., Quirke, P., and Grabsch, H. (2010). The proportion of tumour cells is an independent predictor for survival in colorectal cancer patients. *British journal of cancer*, 102(10):1519–1523.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

Xu, Y., Cui, J., and Puett, D. (2014). *Cancer bioinformatics*. Springer.

Yadav, V. K. and De, S. (2015). An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Briefings in bioinformatics*, 16(2):232–241.

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P. W., Levine, D. A., Carter, S. L., Getz, G., Stemke-Hale,

K., Mills, G. B., and Verhaak, R. G. W. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, 4.

Zhang, L., Conejo-Garcia, J. R., Katsaros, D., Gimotty, P. A., Massobrio, M., Regnani, G., Makrigiannakis, A., Gray, H., Schlienger, K., Liebman, M. N., Rubin, S. C., and Coukos, G. (2003). Intratumoral t cells, recurrence, and survival in epithelial ovarian cancer. *New England Journal of Medicine*, 348(3):203–213. PMID: 12529460.

Zhou, S., Lafferty, J., and Wasserman, L. (2010). Time varying undirected graphs. *Machine Learning*, 80(2):295–319.

Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443.